Application of Machine Learning Algorithms to the Prediction of Water Main Deterioration

Mohammad Amini

A Thesis

In the Department

Of

Building, Civil, and Environmental Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of Master of Applied Science (Civil Engineering) at

Concordia University

Montreal, Quebec, Canada

November 2021

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Mohammad Amini

Entitled: Application of Machine Learning Algorithms to the Prediction of Water Main Deterioration

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

Master of Applied Science (Civil Engineering)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Dr. M. Nik-Bakht Chair

Dr. O. Moselhi Examiner

Dr. M. Nik-Bakht Examiner

Dr. R. Dziedzic Thesis Supervisor(s)

Thesis Supervisor(s

Approved by

Chair of Department or Graduate Program Director Dr. M. Nik-Bakht

December 2, 2021

Dean Dr. Mourad Debbabi

Building, civil, and environmental engineering

Abstract

Application of Machine Learning Algorithms to the Prediction of Water Main Deterioration

Mohammad Amini

Drinking water networks are among the essential infrastructure in cities worldwide. The failure of water mains jeopardizes this essential service and the safety of water users. However, across North America, the failure rate of older water mains has been increasing. The goal of this study is to compare the accuracy and applicability of machine learning algorithms to predict water main deterioration across Canadian water systems. In previous studies, different approaches were applied to only one or a few utilities. Nevertheless, it is valuable to compare results among various networks with different characteristics and levels of data collection. Accordingly, data was collected from thirteen Canadian water utilities, including Barrie, Calgary, Halifax, Kitchener, Markham, Region of Durham, Region of Waterloo, Saskatoon, St. John's, Waterloo, Winnipeg, Victoria, and Vancouver. A variety of factors, including intrinsic, environmental, and operational, were used to develop more reliable predictions and assess the relative importance of each factor. Random forest (RF), artificial neural networks (ANN), extreme gradient boosting (XGBOOST), and logistic regression (LR) were applied to predict the probability of failure. Furthermore, RF, ANN, XGBOOST, and ElasticNet regression models were employed to predict age at first failure and the current rate of failures. Results indicated the superiority of XGBOOST over other models in predicting the probability of failure and the current rate of failure. However, for age at first failure, RF performed better. When datasets were significantly imbalanced, the application of the Synthetic Minority Oversampling Technique (SMOTE) provided more accurate predictions. Because these models provide predictions for every pipe in the network, they can be mapped to facilitate the visualization of deterioration. While models created for one utility cannot be accurately applied to other utilities, the same machine learning algorithms can be quickly and effectively adapted to multiple utilities. Overall, these models support robust and data-driven asset management decision-making.

Acknowledgments

I would like to express my sincere gratitude to Professor Rebecca Dziedzic, my supervisor, for her constant support of my master's study. Without her guidance, this research and writing this thesis would not have been possible, considering the fact that since the beginning of this project, we have been fighting with Covid-19 pandemic. Her precious knowledge, patience, encouragement, and support helped me throughout writing this thesis.

Second, I should thank Concordia University for the financial support and for providing valuable technical tools during the Covid-19 pandemic.

I would also like to express my warmest appreciation to the National Water and Wastewater Benchmarking Initiatives (NWWBI) representatives for providing the datasets and making priceless comments.

Last but not least are my parents for their incomparable love and infinite emotional support. Speaking of support, I am interminably indebted to my lovely wife, Yasaman. Her endless emotional supports have made this path less cumbersome for me. This adventure would not have been even imaginable if not for her.

This thesis shall be devoted to my wife, Yasaman.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AC - Asbestos Cement

CI - Cast Iron

CON - Concrete

CO - Copper

CLPE - Cross-Linked Polyethylene

DI - Ductile Iron

GST - Galvanized Steel

HDPE - High-Density Polyethylene

PB - Polybutylene

PE - Polyethylene

PVC - Polyvinyl Chloride

PVCF - Polyvinyl Chloride Fusible

PVCO - Polyvinyl Chloride Oriented

PCCP - Pre-stressed Concrete Cylinder Pipe

SST - Stainless Steel

ST – Steel

RF – Random Forest

LR – Logistic Regression

XGBOOST – Extreme Gradient Boosting

ANN – Artificial Neural Networks

SMOTE – Synthetic Minority Oversampling Technique

1. INTRODUCTION

In today's fast-paced world, many cities are struggling to manage their aging infrastructure. Evermore stringent practices are being set to manage critical infrastructure, including timely repair and replacement of water systems. Water mains are a vital component of water systems, as they convey drinking water to millions of end-users. Furthermore, as pipes are usually buried underground, their inspection and condition assessment can be cumbersome (Folkman, 2018). Accordingly, various researchers are seeking to develop better measures and methods to assess and predict pipe failure. This would allow utilities to better rehabilitate and replace this essential infrastructure. A deeper understanding of pipe failures could also be applied to asset management practices to keep water systems satisfactory and reduce maintenance costs.

In order to reduce lifecycle costs, water utilities seek to predict the likely time of pipe failure. This enables the identification of the optimal year for pipe replacement or repair. Unexpected water main failures can lead to several challenges for utilities and end-users, such as service interruption, reduced system capacity (Andreou, 1986), and fire-fighting capability. Pipe deterioration on its own can bring about tuberculation, leading to a reduction in system capacity as well (Lei and Saegrov, 1998). Water main failures can also damage other services and nearby properties and incur substantial rehabilitation or replacement expenditures (Shamir and Howard, 1979). Yamijala et al. (2009) reported that another potential consequence is water contamination which puts users' health at risk. Service interruptions and poorer water quality also lead to general customer dissatisfaction.

A study of USA drinking water systems found that most water mains laid during the early to mid 20th century have an average life span of 75 to 100 years. However, if current replacement rates continue, replacing all mains would take approximately 200 years, double the estimated service life of water mains (Folkman, 2018). This period can be reduced by implementing asset management practices. In order to reduce this deficit, some utilities in the mentioned study were reported to have a 125-year replacement strategy (Folkman, 2018).

In 2017, the ASCE report card was prepared and given grade D to drinking water infrastructure in the USA, as opposed to D- in 2009 and an estimation of 240,000 breaks per year in the USA (ASCE, 2017)(ASCE, 2009).

Furthermore, from 2012 to 2018, the total breakage rates surged from 11 to 14 breaks/100 miles/year (Folkman, 2018). Cast Iron and Asbestos Cement pipes account for almost 41% of the installed water mains in the USA and Canada, and many of them are approaching the end of their expected service lives, having experienced an increase of over 40% in breaks rate. Thus, having a robust strategy to maintain these pipes is the primary concern for most utilities (Folkman, 2018).

On the other hand, 59% of pipes in Canada were reported to be less than 40 years old and only 9% above 80 (Canada Infrastructure Report Card, 2019). However, if reinvestment is not

increased in Canada, the condition of core infrastructure may worsen, increasing the cost and risk of service interruption (Canada Infrastructure Report Card, 2016).

Mirza (2007) reported that among the $123 billion estimated total infrastructure backlog, $31 billion was related to water and wastewater networks. Mirza (2007) noted that the total cost to maintain core infrastructure was projected to rise to $400 billion in 2020. The Canada Infrastructure Report Card (2016) found that 29% of portable water infrastructure was in very poor, poor, and fair condition with a cost of $60 billion to replace. The more recent assessment found a similar percentage, of 25% (Canada Infrastructure Report Card, 2019).

1.1 Problem Statement

Previous pipe deterioration prediction studies applied different factors in modeling deterioration, depending on the availability of data for the given case studies and assumed important factors. These applications have focused on only a few water utilities. The broad applicability and accuracy of different predictive models for various utilities are unknown. In order to develop a framework for water main deterioration modeling across Canada, models should be compared under different conditions.

The factors considered by previous studies are also limited, generally focusing on pipe age, material, and diameter. Nevertheless, other intrinsic, environmental, and operational factors have been shown to impact deterioration significantly. Thus, if additional data is available for a particular utility, it may potentially improve deterioration predictions. Accordingly, investigating the importance of predictive attributes cannot only help improve the accuracy and efficiency of deterioration models for utilities with large datasets but also provide insight into additional data other utilities should collect.

1.2 Research Objectives

The primary objective of this study is to compare the accuracy and applicability of machine-learning algorithms applied in predicting water main failure across Canadian water systems. This main objective is achieved through the following specific objectives:

1. Investigate the accuracy and applicability of models to predict probability of failure for different water systems.
2. Investigate the accuracy and applicability of models to predict age at first failure and the current failure rate for different water systems.

Data was collected from thirteen utilities across Canada to meet these objectives, including Barrie, Calgary, Halifax, Kitchener, Markham, Saskatoon, St. John's, Vancouver, Victoria, Region of Durham, Region of Waterloo, Waterloo, and Winnipeg.

2. LITERATURE REVIEW

In recent decades, many studies have been conducted to understand the water mains deterioration process better. As a result, models for water main failure prediction can be broadly classified as physical or statistical. Physical models estimate the residual structural strength of pipes subject to different loads. In contrast, statistical models rely on historical failure data.

The residual structural strength of pipes is affected by many factors such as environmental and operational conditions in addition to manufacturing and installation practices (Rajani and Kleiner, 2001). Physical models estimate the probability of failure by investigating the loads imposed on a water main network and pipes` strength that withstands these loads in the system (Kimutai et al., 2015; Mazumder et al., 2019; Park et al., 2011). Examples of these factors that put pressure on a pipeline include soil pressure, external and internal load (e.g., traffic, earthquake), frost load, and operational water pressure within the network. For instance, the internal load could be a water hammer, and traffic could be an external load. Moreover, these models can be applied by various types of information, such as pipe features, material specifications, pipe age, the severity of corrosion (corrosion condition), different environmental factors such as temperature and rain deficit (Rajani and Kleiner, 2001).

One of the advantages of physical models is that mainly a large number of historical data is not required to develop a predictive model (Wilson et al., 2017), compared to statistical models, in which having large enough data is the primary part of the process. However, the information collected for physical models should include all required details for the analytical processes.

Meanwhile, large diameter pipes are less prone to failure. Therefore, enough historical data is not available. For instance, in this case, physical models are more justifiable where the cost of failure is significant (Kleiner and Rajani, 2001). Physical models are classified into two main categories, deterministic and probabilistic models. The deterministic is defined as a model in which relationships between variables are assumed to be certain (Clair and Sinha, 2012). The deterministic model consists of several models to calculate, for instance, the frost load, the soil-pipe interaction, and the residual structural resistance of water mains (Rajani and Kleiner, 2001).
On the other hand, the probabilistic models mainly focus on the resilience of pipes by predicting the likelihood of failure, integrating a wide range of ambiguities in the physical condition of pipe modeling (Nishiyama and Filion, 2013). Creighton (1994) also believed that the probabilistic models analyze the likelihood of an event occurring. Therefore, this probability related to these incidences is feasible for providing an explanation pertinent to asset failure.

Even though physical models can result in more accurate predictions, they require comprehensive datasets that are not effortlessly available. Therefore, these physical models are generally justified for pipes where failure's cost or broader consequence is considerable (Kleiner and Rajani, 2001). These are usually large pipes (transmission network) with little

redundancy utilized to transfer water from, for example, a reservoir to a local distribution network. Larger pipes are less likely to fail; thus, there exists less accessible historical data. On the other hand, statistical models can be applied with varying levels of data availability and can be used for minor networks (distribution water mains), for which the cost of failure is not considerable (Kleiner and Rajani, 2001). Thus, the focus of this study is on machine learning models created based upon the concept of statistical models.

Owing to the deterioration process's inherent complexity, attempts to predict the failure of water mains focus mainly on statistical models (Lei and Saegrov, 1998). Statistical models use historical failure data to define patterns that are assumed to continue in the future. Statistical models can be either probabilistic or deterministic. In addition, they may estimate the probability of failure, rate of failure, and age at first and subsequent failures (Kleiner and Rajani, 2001; Park et al., 2011). Thus, comprehensive data would, undoubtedly, increase the accuracy of the models (Kleiner and Rajani, 2001).

Deterministic models predict a certain rate of failure or age at failure by fitting different equations to historical data (Kleiner and Rajani, 2001). These models generally require pipes to be partitioned into homogeneous groups with similar characteristics, such as material, size, land under which pipes are laid, soil type, and installation period. This partitioning, however, imposes a challenge on the analysis process. Groups should be large enough and significantly different for the related models to be reliable (Kleiner and Rajani, 2001). Therefore, it is recommended that the size of the dataset be assessed prior to commencing the analytical process.

On the other hand, probabilistic models predict the probability of pipe failure at a specific time in the future or the probability for a pipe to enter into the next stage of deterioration (Andreou, 1986). According to Kleiner and Rajani (2001), the advantage of these models is that they generally remove the need for partitioning, even if partitioning may raise the accuracy of the results. However, the mathematical calculation of these models is highly intricate and requires more expertise than deterministic models (Kleiner and Rajani, 2001). The following sections describe various statistical models in more detail.

2.1 Statistical Deterministic Models

2.1.1 Time Linear Models

One of the pioneering studies in water failure prediction was conducted by Shamir and Howard (1979). They proposed both linear and exponential equations to predict the number of failures based on the pipe age. The linear model, which was not supported with any statistical analysis, is shown in TABLE 2.1. However, the analysis in the study focused on the exponential equation. Walski and Pelliccia (1982) emphasized that the exponential model would be more reliable to relate age to the failure rate. The table lists different time linear models, their accuracies, and more information worth mentioning briefly.

TABLE 2.1 - DETERMINISTIC, TIME–LINEAR MODELS

Authors	Equation and Required Data	Accuracy	Location	Period	Material	Variables in the Equation
Shamir and Howard (1979)	$N(t) = N(t_0)\, A\, (t - t_0)$ Pipe length, Installation date and historical breakage records. For homogeneity: (pipe size, material, soil type, type of failure, and any other criteria that may help to create a better uniform group of pipes)	Not Available	Not available	1961-1975	Not available	t is time in years, $N(t)$ is the rate of failure at time t per year per 1000.ft length, t_0 is an arbitrary base year, $N(t_0)$ is the number of breaks for the base year, and A is the growth rate coefficient *(1/year).*
Clark et al. (1982)	$NY = x_0 + x_1 D - x_2 P - x_3 I - x_4 RES - x_5 LH +$ $NY = 4.13 + 0.338 D - 0.022P - 0.265 0.0983RES - 0.0003LH + 13.28T$ Installation Date, historical breakage record, material, size, operating pressure, soil corrosivity, composition of land overlaying pipes, type of breaks and the pipe vintage which may enhance the accuracy of the model.	$r^2 = 0.23$	Two major cities in North America	1930-1980	Steel and reinforced concrete pipes	NY is the number of years to first repair, D is the pipe diameter (inch), P is operational pressure inside the pipe in psi, I is the percentage of pipe which is covered by industrial area, RES is the percentage of pipe which is covered by residential area, LH is length pipe in interaction with significantly corrosive soil, and T is type of pipe (1 = Metal Pipe, 0 = Reinforced

						Concrete).
Authors	**Equation and Required Data**	**Accuracy**	**Location**	**Date**	**Material**	**Variables in the Equation**
McMullen (1982)	*Age* = x_1*SR* - x_2*pH* - x_3*rd* *Age = 0.028 SR - 6.33 pH - 0.049*r_d Data related to soil characteristics is required. Data could be gathered intermittently which is not expensive. However, gathering data continuously could be highly expensive. Nevertheless, it is recommended to analyze soil around pipes to obtain a model with higher accuracy. It should be noted that water level is not steady, and is related to seasonal precipitation.	r^2 = 0.375,	Des Moines, Iowa		CI pipes	*Age* is time from installation to first failure, *S* is saturated soil resistivity (Ω cm), *pH* is the characteristic of soil, and r_d is redox potential.

Kettler and Goulter (1985)	*N = K_0 * DIAM* *N = 2.002 - 0.0064 DIAM* *N = K_0 * X* *AC pipes (Excluding 23rd winter)* *N = -66.11 + 4.89 X* *CI pipes (Excluding 23rd winter)* *N = -54.29 + 14.29 X* *CI pipes joint failure (Excluding 23rd winter)* *N= -104.6 + 13.79 X* Pipe length, Installation date and historical breakage records. For homogeneity: (pipe size, material, soil type, type of failure, and any other criteria that may help to create a better uniform group of pipes)	r^2 = 0.93 r^2 = 0.884 r^2 = 0.672 r^2 = 0.81	Winnipeg, Manitoba	1950-1959	AC and CI pipes 100-400 mm	*N* is the number of failures, *K_0* is the coefficient, *X* is the number winters, and *D* is the size of pipe.
Jacobs and Karney (1994)	*P = a_0 + a_1 Length + a_2 Age* Pipe length, pipe age (installation data and breakage history) are required. However, in order to have homogenous groups of pipe more data is recommended.	r^2 = 0.704 to 0.937 (All Breaks) r^2 = 0.957 to 0.969 (For independent breaks)	Winnipeg, Manitoba		CI pipes 150 mm	*P* was reciprocal of a probability of a day with no failure, and *a_0*, *a_1* and *a_2* are the coefficient factors.
Yamijala et al. (2009)	$N = \beta 0(D) + \beta 1(AC) + \beta 2(CI) + \beta 3(CSC) + \beta 4(DI) + \beta 5(PVC) + \beta 6(STL) +$	r^2 = 0.12	Texas, USA	2000 - 2005	AC, CI, CSC, DI,	*N* is rate of failures, *D* is the size of pipe in inches; *AC, CI, CSC, DI, PVC* and *STL*

	$\beta 7(L) + \beta 8(Y) + \beta 9(P) + \sum_{j=10}^{20} \beta_j(LUj) + \sum_{k=21}^{25} \beta_k(STK) + \beta 26(TEMP) + \beta 27(RAIN) + \beta 28(SMAX) + \beta 29(MX.MN) + \beta 30(PC1) + \beta 31(PC2) + \beta 32(PC3)$ Size, length, material, type of soil, categorized land use, soil moisture, precipitation amount, average temperature and principle component analysis of soil $N = -0.0027(D) - 0.44(AC) - 0.45(CI) - 0.43(CSC) - 0.46(DI) - 0.45(PVC) + 2.6 * 10^{-5}(L) - 0.00027\ (LU_6) - 0.00032\ (LU_8) - 0.00035\ (LU_{11}) + 0.0018(TEMP) + 3.7 * 10^{-5}(RAIN) + 0.0015(SMAX)$	Holdout sample: zero failure: MSE=0.014 non-zero failure: MSE = 1.12			PVC, and STL	are the binary variable, indicating the type of material; *L* is the length of pipe in feet; *Y* is the installation year of pipe; *P* is the operation pressure inside a pipe in pounds per Sq. Inch; *LU* is the code of land use, categorized into 11 groups; *ST* is the soil type by which pipe is surrounded; *TEMP* is the average of temperature during a 6-month period; *RAIN* is the total amount of precipitation quantified in hundredths of one inch at the nearby airport; *SMAX* is the maximum soil moisture; and *PC1, PC2* and *PC3* are the principal component analysis (PCA) obtained on six different covariates of corrosive soil.

Clark et al. (1982) enhanced Shamir and Howard`s model and proposed a linear model to predict the number of years to the first failure based on several variables listed in TABLE 2.1. The initial assumption was that these variables are independent. However, the model resulted in a low R^2 of 0.23, which indicated that variables might act jointly rather than independently on the failure rate. The authors also found that the interval between subsequent breaks shortens significantly as the number of previous breaks increases.

Jacobs and Karney (1994) utilized linearity intuition with a different approach and tried to predict a time for a pipe without failures, using the length and the age of a pipe. According to previous studies in Winnipeg (Goulter et al., 1993; Goulter and Kazemi, 1988), within which clustering phenomenon was introduced, Jacobs and Karney (1994) proposed a new definition for an independent break. That is, an independent break is a failure that occurs 90 days after the earlier break and/or 20 meters from the previous break. Accordingly, an independent break is the first break that occurs within the break clustering. Goulter et al. (1988), using a clustering method based on time and distance from the previous failures for a group of pipes, mentioned that the powerful inclination for the emergence of failures could be directly pertinent to previous breaks within the vicinity of the new failures, temporally or spatially. This is due to either the weakening process of pipes or the deterioration of bedding around the failure. This weakening and deterioration could rise from surrounding soil affected by the failure, resulting from the maintenance process. Jacob and Karney (1994) introduced a new linear regression and applied it to data from Winnipeg, including 390 Km of 150 mm Cast Iron pipes. The authors categorized water mains into three groups considering their ages (0 to 18, 19 to 30, and >30). Doing so, they achieved uniform data to which their model was applied. The developed linear equation for this model is provided in TABLE 2.1.

They initiated their analysis by applying a linear model to all breaks within the network, resulting in r^2 ranging from 0.704 to 0.937 for different age categories, which indicated that breaks were uniformly distributed within the network. Afterward, according to their definition regarding independent break, they applied the model to those breaks which happened after 90 days and/or within 20 meters from the previous break. The result was significantly appealing since r^2 increased to 0.957-0.969, indicating that the independent breaks distributed across the length of the network uniformly, which certified their assumption about independent breaks. Using age as a factor in the regression analysis, authors improved the quality of their examination for either new pipes and noticeably for older pipes. They realized that the age correlation could be associated with different manufacturing, operation, and installation practices, which were the characteristics of different age groups. Additionally, the authors inferred that these different features could be classified geographically and realized that age would be an appropriate substitute measure that can be collected and managed in GIS (Kleiner and Rajani, 2001).

Kettler and Goulter (1985) analyzed pipe failure considering time increment and increasing pipe diameter in a further study. They noticed that pipe failures decreased as the diameter in CI pipes increased (R^2 = 0.93). The study was implemented on gathered data from four different cities in North America (New York, St. Catharines, Philadelphia, and Winnipeg). However, the

focus of their research was on Winnipeg in Canada, for which the data was desirably comprehensive. Since data for the larger pipe was not readily available, pipes from 100 to 400 mm were chosen for the examination. The main equation utilized for modeling is provided in TABLE 2.1.

This study also mentioned that the larger pipes have thicker walls, leading to less failure rate in those pipes. Kettler and Goulter (1985) also applied the suggested linear equation to the dataset from Winnipeg. However, due to insufficient information, they utilized the number of winters after 1959 (from 16th winter in 1975 to 23rd winter in 1983) to define the sample for their examination. Since the number of failures for the 23rd winter was considerably low, they decided to implement their analysis including and excluding that specific winter and considered the 23rd winter an outlier in parts of their analysis. Related equations are provided in TABLE 2.1. This model was relatively straightforward to be applied. However, the authors did not provide any validation for the model.

McMullen (1982) suggested a linear relationship between pipe failure and soil specifications and proposed a model applied to the water distribution network of Des Moines, Iowa. The researchers in the examination team realized that 94% of failures were attributed to soil with less than 2000 Ω cm saturated resistivity, and corrosion was found as the primary type of failure in the system. According to the low to moderate $r^{2,}$ which was 0.375, it was certified that soil was the major contributing factor to pipe failures (Kleiner and Rajani, 2001), leading to expected service life reduction by 28 years for every 1000 Ω cm decrease in soil resistivity. Moreover, similar to that of Clark et al.(1982), the low value of r^2 implies that the factors utilized in the analysis may not act independently; therefore, they should be employed dependently and multiplicatively. The proposed equation and the equation with suggested coefficients are provided in TABLE 2.1.

Yamijala et al.(2009) incorporated a wide variety of features to the basic model introduced by Kettler and Goulter (1985). The analysis process started with 33 different variables. However, they finally ended up with fewer numbers of covariates with a significant p-value around 0.05. Implementing many analytical iterations, the authors ultimately obtained the equations provided in TABLE 2.1. The accuracy of this model was relatively low, with an r^2 of 0.12. However, this does not indicate that whether the model would fit the dataset or not. That is, a more analytical process is required to find out about the accuracy rate of this model. Finally, this model depicted a low accuracy for non-zero failure among the pipes.

2.1.2 Time Exponential Models

Time exponential is a non-linear model proposed for the first time by Shamir and Howard (1979). The authors believed that the historical data must be investigated and utilized to foresee how the rate of failures in existing pipes will change in the future. It was reckoned that there is a positive correlation between pipe failure and the aging process and realized that the rate of failure increases exponentially as a pipe is aging. In this model, regression analysis was

used, which was supposed to predict the rate of breaks by pertaining pipe failures to the exponent of its age. The equation for this model is provided in (TABLE 2.2).

TABLE 2.2 - DETERMINISTIC, TIME – EXPONENTIAL MODELS

Authors	Equation and Required Data	Accuracy	Location	Date	Material	Variables in the Equation
Shamir and Howard (1979)	$N(t) = N(t_0)\, e^{A\,(t-t_0)}$ Pipe length, Installation date and historical breakage records. For homogeneity: (pipe size, material, soil type, type of failure, and any other criteria that may help to create a better uniform group of pipes)	Not available	Not available	1961 - 1975	Not available	t is time in years, $N(t)$ is the rate of failure at time t per year per 1000.ft length, t_0 is an arbitrary value which could be either the installation year or a year from which first information has been extracted, $N(t_0)$ is the number of breaks for the base year, and A is the growth rate coefficient.
Walski & Pelliccia (1982)	$N(t) = C_1\, C_2\, a\, e^{A\,(t-t_0)}$ Data required: Same as Shamir and Howard (1979) Pit CI: $N(t) = 0.02577\, e^{0.0207\,(t-t_0)}$ Sand spun CI: $N(t) = 0.0627\, e^{0.0137\,(t-t_0)}$	Not available	Binghamton, NY	Not availabl e	Pit CI and Sandspun CI 100-600 mm	t is time in years, t_0 is the installation year, $N(t)$ is the rate of failure at age t (break/mile/year), a is the coefficient factor calculated by regression analysis, A is the growth factor coefficient (1/year), C_1 is the added, factor for previous breaks, and C_2 is the added, factor for size. These two factors are created based on the available information
Clark et al. (1982)	$REP = y_1 * (e^{y_2})^{T} (e^{y_3})^{PRD} (e^{y_4})^{A} (e^{y_5})^{DEV} (SL^{y_6})(SH^{y_7})$ $REP = 0.1721 * (e^{0.7197})^{T} (e^{0.0044})^{PRD} (e^{0.865})^{A} (e^{0.0121})^{DEV} (SL^{0.014})(SH^{0.069})$	$r^2 = 0.47$	Two major cities in North America	1930-1980	Steel and reinforced concrete pipes	T is type of pipe (Metallic or Concrete) A is the age of pipe from the first failure, REP is the quantity of repair, PRD is differential pressure in psi, DEV is the percentage of the area above pipe in low corrosive and moderate

						corrosive soil, *SL* is the surface area of pipe in low corrosive soil, and *SH* is surface area of pipe in highly corrosive soil.
Y. Kleiner & Rajani (2002)	$N(x_t) = N(x_{t_0})\, e^{a\, x_t^T}$	r^2 = 0.619 to 0.793	Ottawa – Carlton, Ontario	1973-1998	CI and DI pipes	X_t is the vector of Time-Dependent variables which exist in time t; $N(x_t)$ is the breaks rate as a result of vector of X_t, a is corresponding coefficients related to covariate x, x_{t_0} is vector of baseline in the year t_0 which is the reference year, and $N(x_{t_0})$ is the rate of failure for the base year.
Yamijala et al. (2009)	$Y = \beta_0\, e^{(\beta_1 (TIME) + \beta_2 (INSTYR))}$ $Y = 0.093\, e^{(0.47 (TIME) + 2.7*10\text{-}5 (INSTYR))}$	Holdout Sample : zero failure: MSE = 0.03 non-zero failure: MSE = 0.78	Texas, USA	2000 - 2005	AC, CI, CSC, DI, PVC, and STL	β_0 is coefficient factor calculated by regression analysis, β_1 is coefficient factor of time elapsed from the first break, β_2 is coefficient factor of year in which pipe was installed, *Y* is rate of failure, *INSTYR* is the installation year, and *TIME* is the elapsed time from the last break.

Shamir and Howard (1979) assumed that $N(t_0) \neq 0$, meaning that even a newly installed pipe has a negligible number of breaks that may not be overlooked.

In this study, for various groups of pipes, a range of 0.01-0.15 for coefficient *A* was proposed. They also recommended that to utilize this model to achieve a reliable result, the population of pipes should be partitioned into homogenous groups.

There are, however, some critiques to this exponential model. Although the model was relatively straightforward to be applied to data, it required significant attention to partition data into homogenous groups. Besides, the authors did not provide adequate information about the location of the investigation. Meanwhile, the quality and method of analysis were not mentioned clearly. They also presumably considered failures to be distributed within the network uniformly. Consequently, many authors questioned their model in further studies (Kleiner and Rajani, 2001).

Several years later, Walski and Pelliccia (1982) enhanced the model. Comparing both linear and exponential models, they realized that the exponential model best fit the data. The authors added up two additional factors to the primary model in order to increase the result accuracy. The model was based on the observation conducted by the US Army Corps of Engineers in Binghamton, NY (Walski and Pelliccia, 1982).

The first factor (C_1) considered the previous historical breaks of a pipe. If a pipe had one break prior to analyzing, it was more likely to fail again in the future. The second factor (C_2) applied was related to the pipe with a larger diameter in Pit Cast Iron pipes. They also changed the arbitrary t_0 in the main equation of Shamir and Howard (1982) to the installation year; thus, $(t-t_0)$ became the pipe age. The proposed equations by Walski and Pelliccia are provided in (TABLE 2.2). The low values of (A) indicated that the failure rate does not necessarily surge significantly while a pipe is aging. Furthermore, the impact of temperature was analyzed in this study and realized that incorporating the coldest months of the year may be a significant factor in predicting the water mains failures. However, this factor may not be feasible to be included in the analysis since the severity of winter is somehow unforeseeable.

According to Walski and Pelliccia (1982), this model may be considered fundamental to predict pipes failures, albeit it did not employ pipes with a higher rate of failures which usually happen within the wear-out phase of the bathtub curve (Figure 2.1). More importantly, the authors did not mention whether the additional factors (C_1 and C_2) enhanced prediction quality and how far (Kleiner and Rajani, 2001). Finally, C_1 and C_2 were chosen randomly, which may not be considered as a logical selection statistically.

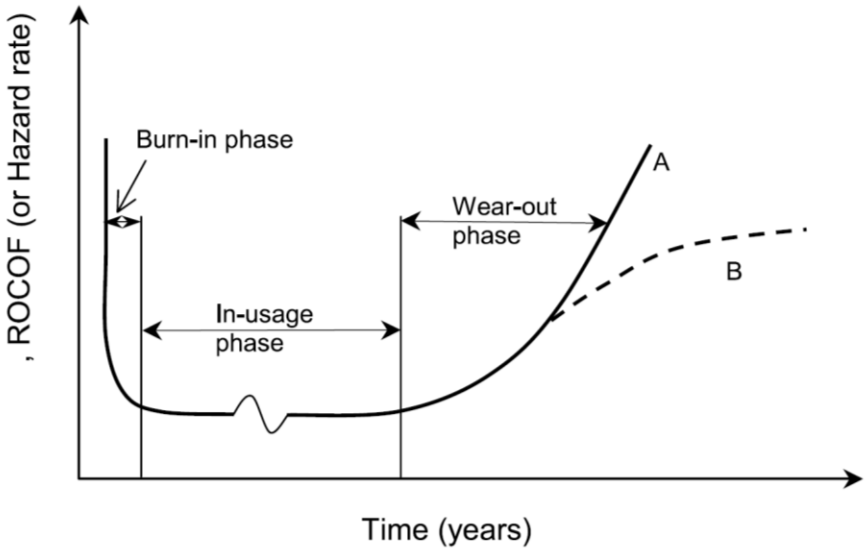


FIGURE 2.1 – THE BATHTUB CURVE OF LIFE CYCLE OF A BURIED PIPE (KLEINER AND RAJANI, 2001)

In a later study, Yamijala et al. (2009) modified the exponential model of Shamir and Howard (1979) and introduced a manipulated model to predict the rate of breaks. The analysis for this model was conducted upon data received from a utility in Texas, USA (TABLE 2.2). Analyzing the result indicated that the number of failures would increase as the longer time elapses from the last failure. The result also depicted that the non-linear model is more reliable and accurate than the linear model. However, the overall result indicated a non-strong fit of the model to the data.

In another endeavor to enhance the Shamir and Howard (1979) model, Clark et al. (1982) proposed a new model and altered the exponential model to predict the frequency of repairs after the first break for an individual pipe (TABLE 2.2). The r^2 was calculated moderately as 0.47 for the exponential model. This equation should be applied to each pipe individually to predict the number of failures after the first failure. The moderate r^2 implies that the model should be analyzed in order to find out whether it is suitable to be used or not. The model was likely to improve if there were more factors available to be included.

Kleiner and Rajani (2002) mentioned that no statistical models utilized the time-dependent factors except the age of pipe and number of previous breaks. Therefore, they introduced a multi-variate time-exponential model considering time-dependent factors which may affect pipe deterioration. Authors believed that it is true that a pipe may deteriorate steadily and monotonously, yet some environmental and operational factors may affect this process depending on time, and they could be ephemeral. Some of those factors could be the impact of aging, temperature, moisture within the soil, aggregated length of replacement, and cumulative length of retrofitted water mains. A worthwhile predecessor to this model is to partition water mains into homogenous groups that respond similarly to the deterioration process. The authors applied this model to data from Ottawa-Carlton, Ontario. They mainly

focused on cast iron and ductile iron pipe, for which data from 1973 to 1998 was available. The manipulated format of the exponential model suggested by Kleiner and Rajani (2002) can be seen in (TABLE 2.2)

2.2 Statistical Probabilistic Models

2.2.1 Weibull Distribution

The well-known Weibull distribution is an adaptable approach that can be employed to describe failure datasets. Also, Weibull is the merely parametric regression approach that has both accelerated failure representation and proportional hazards presentation (Røstum, 2000).

The WPHM or Weibull Proportional Hazards Model is an approach that is used to model times to failure distribution and links a variety of explanatory attributes to this interval time (Le Gat and Eisenbeis, 2000; Kalbfleisch and Prentice, 1980). These covariates could affect the failure rate as well as the number of years to failure. A critical point in this model is that the time to failure could be either right-censored or observed failure.

2.2.2 Poisson

The Poisson distribution is one of the most well-known statistical models that explain the probable number of events (number of failures for a specific pipe) or objects in a particular volume in a specific interval of time (Motulsky, 2010). Some criteria should be taken into consideration while implementing Poisson distribution (Jarosz, 2021):

1- Events should be countable in positive format (0, 1, 2, …)

2- The occurring events are typically independent and random

3- Each object or event is counted only one time

4- The average frequency of occurrence is predefined and known

2.2.3 Proportional Hazards Model

The proportional Hazards Model was proposed for the first time by Cox in 1972. For the group of individuals, the survival time (time to failure or the time to loss) is observed. For the individuals who have been censored, it is merely known that the time to failure is higher than the censored time.

In many studies before, several regression analyses were proposed to define an appropriate relationship between the survivor time t and an array of covariates Z, among which PHM (proportional hazards model) introduced Cox (1972) found to be more feasible and have been applied by many researchers (Andreou, 1986, Andreou et al. 1987a, Jeffrey, 1985, Lei and Saegrov, 1998, Vanrenterghem-Raven et al. 2003). One of the most significant upsides of survival analysis is that it considers different pipes with one or more failures during the examination and those pipes that have not been failed during the analysis (Mailhot et al., 2000). More importantly, being dynamic, the PHM model may be used by one to update the probability of failures for a pipe towards future time after each break. The hazard function is provided in TABLE 2.3.

According to Cox (1972), the baseline hazard function which is $h_0(t)$, can be deciphered as a time-dependent aging element. Whereas the other covariates stand for operational and environmental factors, acting on the water pipes to decline or even increase the probability of failure.

For the first time, the methodology of PHM application to predict the probability of failure for a particular pipe at a specific time in the future within the water main networks was introduced by Jeffrey (1985). This study was conducted upon data from New Haven, Connecticut in The USA, including pipes with the number of failures less than four, as well as a pipe with a large diameter since it was believed that these kinds of pipes have more effects on managerial decisions than those of smaller pipes. The author utilized various regression methods to find out covariates Z that might significantly contribute to pipe failure rates. The covariates found by the examinations are listed in TABLE 2.3. The arbitrary baseline hazard function $h_0(t)$ was computed using the polynomial regression analysis.

TABLE 2.3 - PROBABILISTIC MODELS – PROPORTIONAL HAZARDS MODEL (PHM)

Authors	Equation and Required Data	Location	Date	Material	Variables in the Equation
Jeffrey (1985)	$h(t,Z) = h_0(t)\, e^{b^T Z}$ Natural log of pipe length, pressure in the system, age, the number of previous breaks, vintage of pipe, the proportion of low land development. Baseline hazard function: $h_0(t) = 10^{-4}(2 - 0.1\,t + 0.002\,t^2)$	New Haven, Connecticut	1930-1985	Different Pipes in the system	*t* is the time, *h(t, Z)* is the hazard function, $h_0(t)$ is the arbitrary baseline hazard function, *Z* is the covariates vector which acts multiplicatively on the hazard function, and *b* is the coefficients vector. *t*
Andreou S. A. (1986) Andreou et al. (1987a)	First Stage: $h(t,Z) = h_0(t)\, e^{b^T Z}$ Natural log of pipe length, pressure in the system, age, the number of previous breaks, vintage of pipe, the proportion of low land development. Baseline hazard function: $h_0(t) = 10^{-4}(2 - 0.1\,t + 0.002\,t^2)$ Second Stage: $\lambda = \exp(b^T z) + e$ $P(x) = \frac{(\lambda t)e^{-\lambda t}}{x!} \qquad x = 1,2,3, \ldots.$	New Haven, Connecticut and Cincinnati, Ohio	1855 - 1985	Different Pipes in the system Pipe with diameter equal or greater than 8 inches	PHM variables: Same as Lisa A. Jeffrey (1985) Poisson type model: λ is annual rate of failure, *b* is vector of coefficients, *z* is vector of covariates, and *e* is model error *p(x)* is the probability of x breaks, *t* is time, and *x* is the number of breaks.
Le Gat and Eisenbeis (2000)	$\ln T = x^T \beta + \sigma W + \mu$	Data from Charente-Maritime and Lausanne	Different Periods	AC,CI, Steel and PVC	*T* is a pipe lifetime, *x* is a vector of explanatory variables; β is a vector of coefficient related to each explanatory variables which can be estimated by max likelihood using different methods (e.g. Newton-Raphson algorithm); σ is an unknown parameter than can be estimated

					by maximum likelihood; W is a random variable (Weibull distribution); and μ is a constant value;
Vanrenterghem-Raven et al. (2003)	$\ln T = x^T\beta + \sigma W + \mu$ length, diameter, material, age, traffic, soil, subway, location in the street, presence of ancient water zones	New York, USA	1982-2002	CI, Lined CI, DI and Steel 4 – 72 inches	T is a pipe lifetime, x is a vector of explanatory variables; β is a vector of coefficient related to each explanatory variables which can be estimated by max likelihood using different methods (e.g. Newton-Raphson algorithm); σ is an unknown parameter than can be estimated by maximum likelihood; W is a random variable (Weibull distribution); and μ is a constant value;

Given is the $h_0(t)$ utilized in Jeffrey, (1985) research:

$$h_0(t) = (2 * 10^{-4}) - (10^{-5}t) + (2 * 10^{-7} t^2) \quad (1)$$

Where:
t is the time for a pipe to survive since the installation for a new pipe, and since the last break for a pipe with the previous break;

The $h_0(t)$ in their analysis does not correspond to the life cycle bathtub curve (ROCOF); instead, it demonstrates the immediate hazard (probability of next break) after installation or after the last break for a new pipe or a pipe with previous failures, respectively. The break hazard had a minimum of t=28 within the Jeffrey (1985) examination, indicating that a 28-year old pipe had a minimum risk of failure. Therefore, according to their analysis, a pipe's probability of failure decreases throughout the aging period after installation until 28 years. Afterward, the pipe initiates the deterioration process, and the probability of failure increases. Likewise, this process would happen for a non-new pipe, increasing the likelihood of failure 28 years after the previous failure.

This model then was enhanced to foresee the probability of failure in a water network and realized that a two-stage analysis would be required (Andreou et al., 1987b, 1987a, Andreou, 1986). That is, in order to accomplish a model with higher accuracy, defining different stages in which a pipe may fail would be noticeably significant. The early stage of failure included fewer breaks, and the late stage of deterioration comprised multiple and frequent failures. Therefore, the authors suggested that the proportional hazards model (PHM) could be utilized for the early stage and the Poisson type model for the late stage.

The authors stated that the failure patterns among different water systems are not consistent; thus, analyzing each pipe is of significant importance, helping to make a better decision to conduct maintenance practices. In addition, pipes with a larger diameter are less prone to failure. Therefore, they may be considered to become appropriate candidates to implement rehabilitation strategies. On the contrary, rehabilitation preferences for smaller pipes are not economically justifiable. Thus, the authors focused on the cast iron pipes with a diameter equal to or greater than 8 inches.

Initial statistical analysis revealed that as a pipe was experiencing the first break, the time to subsequent break shortened. However, from the third break, there seemed to be no clear pattern. Several analyses inferred that previous breaks might not be considered an important index to predict the subsequent failures after the third break. Therefore, the third break was selected as a cut-off point for the analysis. Moreover, the need for another model to examine data for two networks was identified. Finally, the authors mentioned that stage classification is highly site-specific and may not be applied to other utilities. However, it gives an insight into

the importance of deterioration stage classification for further study in the future (Andreou et al., 1987b). Significantly noteworthy is that gathered data in terms of break records and operational and environmental factors require accurate interpretation to increase the precision of analysis (Andreou et al., 1987a).

Furthermore, Andreou et al. (1987) realized that quantifying the probability of failure is significantly important. Accordingly, the very compelling model to predict the probability of failure was considered to be the survivor function for each pipe individually, which predicts the likelihood of surviving toward the future time, followed by hazard function, which leads to the instantaneous probability of breaks for a specific pipe at any given time.

After analyzing the experimental process, some variables appeared to be statistically significant, which can be found in TABLE 2.3. Finally, the baseline hazard function proposed by the authors was similar to that of Jeffrey (1985).

By looking at the baseline hazard function, it can be inferred that the hazard of failure would decrease initially after pipe installation or after the last failure. Using the derivative form of the baseline hazard function, the authors realized that the minimum hazard is at age 28. The derivative form is as follow:

$$\frac{d\ h_0(t)}{dt} = 0 \qquad (2)$$

The bathtub shape of the baseline hazard equation in Andreou et al. (1987a) and Andreou (1986) was intriguing, indicating that a pipe would be in a desirable condition early after installation if it does not have any failure caused by either environmental or operational incidence. However, it also depicted that age would contribute to failure many years after installation when external and internal corrosion affects the pipe deterioration.

In order to predict the probability of failure during the fast-breaking stage, the third and sixth breaks were defined as cut-offs. Andreou et al. (1987) utilized an exponential regression model to estimate the annual failure rate, denoting by λ, provided in TABLE 2.3. Afterward, this λ was added to the Homogeneous Poisson Model to predict the probability of failure. Using the Poisson equation, Andreou et al. (1987) predicted x failure within the network throughout the time *t*. This Homogenous Poisson Model assumes that the rate of failure typically remains unchanged in a time interval. This assumption then was challenged by Goulter and Kazemi (1988). Andreou et al. (1987) also analyzed the effect of cleaning and lining on pipe failures. However, these two factors were found statistically inconsequential.

Conducting the PHM analysis in combination with the Poisson Model, the authors perceived that it would be beneficial to identify a pipe with a higher hazard rate. Thus, an appropriate maintenance strategy may be considered for that specific pipe (Andreou et al., 1987a).

The authors also noted that the pipe age as a contentious feature might have a different impact on pipe failure. That is, not all pipes installed in the previous era are prone to more failure than pipes installed in the more recent periods; thus, failure may be affected by the existence of various factors. Amidst different influential factors which may cause an increase in the rate of failure, higher internal pressure, and land development were found to be significantly important.

Eisenbeis (1994) and Bremond (1997) reported applying the proportional hazards model to data from Bordeaux, France. The Weibull model was used to create the baseline hazard function in their studies. Using 33-year data of Bordeaux, the authors reported a reasonable estimation of the failure rate for 11 years. Nonetheless, whether they performed the two-stage analysis proposed by Andreou et al. (1987) or not has remained unclear.

Lei and Saegrov (1998) also applied the proportional hazard model and the accelerated lifetime model to data from Trondheim municipality, Norway. There did not exist any indication of the quality of the proportional hazards model in this research. However, explicitly can be noticed that pipe material was used to stratify the data rather than as an explanatory variable.

Le Gat and Eisenbeis (2000) applied Weibull Proportional Hazards Model (WPHM) to data from two different companies with long and short maintenance records data. WPHM is an accelerated lifetime model which assumes the time to T (failure) is pertinent to p covariates X with a linear function (TABLE 2.3). These X variables could be pipe characteristics such as size, material, and length, or environmental and operational factors. The number of previous failures, as an important factor, was suggested to be used as a covariate or as a means to stratify the existent data. Le Gat and Eisenbeis (2000) also utilized the Monte Carlo simulation method to predict the rate of failures. The result of this study indicated that WPHM could be applied to networks with long maintenance records and networks with short records. It was also mentioned that a universal method for all networks is less likely to be achieved due to different factors that uniquely affect each network. Moreover, Le Gat and Eisenbeis (2000) noted that this method might be associated with the Geographical Information System (GIS) to accomplish better rehabilitation strategies.

In a later study by Vanrenterghem-Raven et al. (2003), PHM was examined to identify the applicability of this model for the large urban area. The study was conducted upon data from New York, USA, and certified that there is no obstacle to using this model for large cities. It was also noted that material stratification could improve the accuracy of the model.
Park (2004) applied the PHM to investigate the alteration of the hazard function between subsequent breaks. It was suggested that the point that hazard function is changed from an exponential format to a consistent format might be ascertained as a new stage of deterioration for a pipe. Thus, that stage could be examined differently. Park et al. (2008) also applied the PHM and realized that the groups examined in their study followed Weibull distribution. Furthermore, it was cited that the hazard rate of the first break increases as time elapses from the installation, and the time between subsequent breaks declines while the number of breaks increases. In this study, the effect of land development under which pipe is laid and the length

of pipe was found to be significant factors. Finally, p) investigated PHM, applying to the data from a utility situated in the USA. The main focus of the study was on the CI pipe with the size of 150 mm, which presented a significant proportion of different pipes in that network. In addition, the authors included the time-dependent impact of covariates on the hazard rate of pipes. Interestingly, they found the hazard rate pattern similar to the bathtub curve, especially from the third break to the seventh break. However, it was suggested that if a utility opts for using this method, it should inevitably monitor and update data regularly.

Kimutai et al. (2015) compared three widely used probabilistic models, Cox-PHM, WPHM, and Poisson Process model, to find out the superior fitness of these models to data from Calgary, Alberta. The authors examined different groups of materials, dominantly CI, DI, and PVC pipes. For all materials, WPHM and Poisson Process performed better compared to Cox-PHM. However, comparing WPHM and Poisson Process, authors realized that WPHM best fit the metallic pipes. Whereas, Poisson Process appeared to be a better predictive model for PVC pipes. Cox-PHM was found unsuitable for pipes entering the fast-breaking stage, as was mentioned before (Andreou et al., 1987a and Andreou, 1986); consequently, it may be used for younger systems. On the contrary, Poisson Process is inferred to be an appropriate model for pipes entering the fast-breaking stage of their deterioration process. The authors also mentioned that since the utilities worldwide are to change their pipes continuously in terms of material, combining with the older pipes, it is highly suggested to use different models to evaluate the condition of their systems.

In one study Bayesian Model Averaging (BMA) was employed to predict the rate of failure in Kelowna, BC, and Greater Vernon Water, BC, Canada (Kabir et al., 2015). Results of this study indicated that the performance of BMA was better than classical linear regression whenever the information of pipe failures is limited. Pipe age and length were found to be the most contributing factors where data is available limitedly. The authors proposed strengthening the model by integrating it with GIS tools to identify hot spot zones and provide more effective plans for the utilities. Moreover, rain deficit and freezing index, for instance, could be employed to improve the model's reliability. The authors updated the model by proposing another Bayesian Weibull Proportional Hazard Model (BWPHM)(Kabir et al., 2016). BMA was conducted to choose the most critical input variables. Whereas BWPHM was a method to develop the survival shape for Cast Iron and Ductile Iron pipes, applying to 57 years of recorded data from Calgary, Canada. This study indicated that the water main predictive models could be effectively enhanced with an updated Bayesian model.

Snider and McBean (2018) provided an in-depth comparison of one machine learning model (XGBOOST) to Weibull Proportional Hazard survival analysis. The result of this study revealed that the machine learning approach consistently forecasted earlier times to break than actual events. However, survival analysis over-forecasted these incidences. This difference was believed to be related to not incorporating censored data by the machine-learning model (removed information for the training set). Removing this information would lead to training the model ineffectively. The authors suggested that for an immediate maintenance process, machine learning models may be sufficient. However, for finding a more effective long-term

pattern, survival analysis should be incorporated within the analytics, which means that censored data (removed data) should be considered for water main failure prediction.

2.3 Machine Learning Algorithms

Machine learning is an element of artificial intelligence, even though it typically tries to solve problems based on historical records (Rebala et al., 2019; Swamynathan, 2019; Verdhan, 2020). The primary types of machine learning models can be categorized as follows (Rebala et al., 2019):

- Supervised Learning
- Unsupervised Learning
- Semi-supervised Learning
- Reinforcement Learning

In supervised learning, the target is given in the data. For example, in the prediction of pipe failure, it is clear whether a pipe is broken or not. This type of machine learning is the most widely used method (Verdhan, 2020). Generally, there are two types of widely used supervised learning algorithms (Swamynathan, 2019), regression and classification.

Regression predicts a continuous number related to the input variables. For instance, in the water industry domain, the rate of failure or age at failure could be predicted through regression.

Classification forecasts a class (e.g., yes/no) and the corresponding probability of a given class. For example, with this method, the status of a pipe (Broken/Non-Broken pipes) and the corresponding probability of pipe failure at any given time in the future can be predicted. Overall, this method is able to classify data points into distinct classes (Rebala et al., 2019).

In unsupervised learning, the target is not available in the data (Swamynathan, 2019). Instead, unsupervised learning aims to find a pattern for input variables to understand similarity and dissimilarity within unlabelled datasets. For example, this could be finding groups of homogenous pipes to have a better prediction. Since this method does not require consultation with domain experts prior to modeling, it is called unsupervised learning (Swamynathan, 2019).

Semi-supervised falls between supervised and unsupervised learning. In this case, only a few target instances are labeled clearly in the dataset(Rebala et al., 2019). Reinforcement learning improves models iteratively as more data is available and is growing in popularity. This method is mainly used for real-time decisions, games, learning tasks, robot navigation, and skill acquisition.

One issue that must be addressed in applying machine learning models is the availability of imbalanced data. Imbalanced data is characterized by a minority class with peculiar or prominent information and a majority class with standard information (Mena and Gonzalez,

2006). For example, the minority class in the water domain could be the number of failed pipes, and the majority, the number of pipes that have not failed. When the ratio between the two classes surpasses 1:10, they are considered imbalanced. Standard classification algorithms do not perform well when the samples within one class are markedly outnumbered by those in the other class (Wang et al., 2013). Approaches such as Synthetic Minority Oversampling Technique (SMOTE) can be employed to artificially "re-balance" the datasets. This method is explained in more detail in chapter three.

2.3.1 Artificial Neural Networks

Artificial Neural Networks (ANNs) have various applications, and it has been widely used in recent decades for water main failure prediction (Asnaashari and Shahrour, 2007; Moselhi and Shehab-Eldeen, 2000; Sattar et al., 2019; Tavakoli, 2018). In addition, ANNs are networks of multiple neuron layers and can be used for classification and regression tasks (Sharma et al., 2020; Tavakoli, 2018).

ANNs were inspired by the human brain structure, where a complex network of neurons swap information through synapses (Kerwin et al., 2020). There are different types of ANNs; however, the multi-layer perceptron (MLP) is the most popular format found in the studies in accordance with pipe deterioration prediction. Typically, MLP consists of a different number of layers (basically three layers), within which there are some nodes or units called neurons (Giraldo-González and Rodríguez, 2020; Kerwin et al., 2020; Tabesh et al., 2009). Mathematically, a neuron (Perceptron or Node) may be a nonlinear function, which makes ANNs a fully intricate nonlinear system (Tabesh et al., 2009). A three-layer MLP consists of one input layer, one hidden layer (typically), and finally, one output layer. Each layer could be a combination of one or more interconnected neurons, which may facilitate the process of weighting input variables and converting these variables to an output variable using an activation function (Giraldo-González and Rodríguez, 2020; Kerwin et al., 2020). There are a number of activation functions that can be applied for giving specific weights to each attribute. Sigmoid function, tanh, and are among the most common activation functions (Sharma et al., 2020). Yet, there still are many other functions that are explained in more detail in the next section of this study. Sharma et al. (2020) reported that should an activation function not be used in the ANNs, the output would be a simple linear regression. The complexity of linear regressions is limited, and they cannot recognize intricate relationships between different input variables (Sharma et al., 2020), meaning that an activation function can handle more complex relationships within the Neural Networks. In ANNs, the signals are transferred from perceptron i to perceptron j, and the output variable can be defined as the given equation (Figure 2.2).

$$y_i = f\left(\sum_{j=1}^{N} w_{ij} * x_i + b_i\right) \qquad \text{where } i, j = 1, 2, \ldots, N$$

FIGURE 2.2 – OUTPUT EQUATION OF ARTIFICIAL NEURAL NETWORKS

Where:

Xi are the input signals or the explanatory variables;

Wij are the synaptic wrights; and

F is the neuron activation function

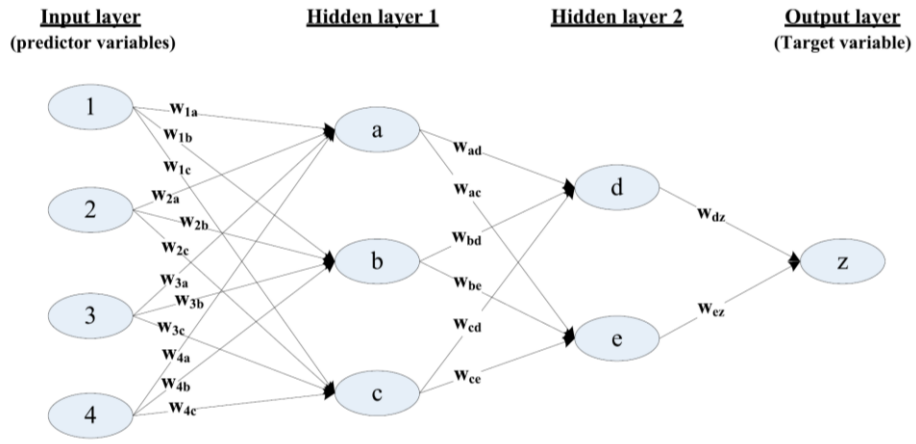


FIGURE 2.3 – NEURAL NETWORK MODEL WITH TWO HIDDEN LAYERS

In order to detect and classify sewer pipe defects, Moselhi and Shehab-Eldeen (2000) applied artificial neural networks (ANN) to data provided by a specific contractor in Montreal, Canada. This study aimed to stratify four types of failures in the underground sewer pipes: joint displacements, spalling, reduction of cross-sectional area, and cracks. The backpropagation algorithm was utilized to learn the model. This algorithm was suggested to be an appropriate method for classification problems. In this study, the result was fascinating, indicating 98% accuracy in defects detection. The developed ANN was able to detect 214 out of 218 defects properly.

Several years later, Ahn et al. (2005) applied ANN model to Seoul City, South Korea data. They investigated the correlation of pipes failures in the service pipes and water mains, considering different factors, which were believed to be highly influential. Recorded failures and changing water and soil temperature were taken into consideration in their analysis, which is related to seasonal change. In addition, they utilized historical records to predict pipe failures to decline the operational cost and improve the reliability of the water distribution network. Researchers utilized data, from 1995 to 2004, on soil and water temperature combined with historical failure records so as to foresee the rate of failures. Authors discovered that the rate of failures for pipes surged following water and soil temperature alteration in fall and spring. The compelling part of their study was that water mains were affected higher by water and soil temperature than service pipes. It was also noted that valves and fittings are more prone to failure due to temperature varying. The ANN model represented a satisfactory result due to the low rate of Mean Absolute Error metric (MAE), except for the time in which the failures rate increased or decreased substantially where the model was less sensitive to accurately

predicting the rate of failure. It was, therefore, suggested that incorporating more factors and the enhanced use of the ANN model would bring about desirable predictive results for utilities.

In another study conducted by Najafi and Kulandaivel (2005), the condition rate of sewer pipes was examined. Several features were considered in this research, such as size, pipe material, slope, type of sewer, depth, age, and section length. The BPNN (backpropagation neural networks) algorithm was applied to data from Atlanta, USA. The RMSE error for training and test sets were calculated as 0.1868 and 0.1792, respectively. It was noted that the result of the examination was insufficient for a desirable statistical analysis. However, the ANN indicated its capability in capturing the sophisticated correlation between inputs, which was a manifestation of the condition state of sewer pipes. Therefore, this model proved an appropriate learning tendency according to the available data. The authors, nonetheless, mentioned that more data availability would lead to this model's higher accuracy and reliability, leading to the prediction of actual pipe conditions. In this study, noises, outliers, and missing values significantly affected the model accuracy.

Furthermore, Al-Barqawi and Zayed (2006) utilized the supervised ANN model (backpropagation algorithm) to evaluate and predict the condition rate of water mains, even before the additional examination. They believed that condition assessment of water mains is among the most challenging obstacles that most utilities are confronting worldwide. The collection of data from three various municipalities was examined. Moncton (New Brunswick, Canada), London (Ontario, Canada), Longueuil (Quebec, Canada) were the case studies for this research. Their datasets, including pipes materials, installation year, size, number of breaks, soil type, C factor (Hazen–Williams factor), depth of the pipe, and the surface type, were used to model the training and test data set. These are the combination of intrinsic, environmental, and operational factors affecting pipe deterioration. In order to evaluate the accuracy of the model, several error terms were taken into account. These evaluation terms indicated a desirable result of the examination.

Furthermore, the authors provided a scale rating condition proposal that can be used by different utilities according to their criteria, based on this research. This study has proven the importance of previous breakage rate and age of mains as the most contributing factors to assess the condition of water main with weights of 30.17% and 13.56%, respectively. Compelling is the part where the authors provided different equations for computing the breakage rate of AC, CI, and DI pipes by which practitioners would evaluate the condition rate of any specific pipe. Finally, comparing breakage rate and condition score, the researchers realized an inverse correlation between breakage rate and condition rate. This study indicated the robustness of the ANN model and its feasibility to ascertain the condition state of water mains.

Al-Barqawi and Zayed (2008), in further study, tried to design a comprehensive model to evaluate the condition and performance of water mains. The same data as the previous study was used (Al-Barqawi and Zayed, 2006). The authors created an integrated model using the analytic hierarchy process (AHP) and artificial neural networks (ANN). One specific infrastructure management tool was developed based on the produced integrated model. It

should be noted that the accuracy of the model was reported at 98.51%, and also age was found to have a significant impact on the condition rate of the pipe, followed by material and breakage rate. Nevertheless, the accuracy of the result indicated that the model proposed by the authors is efficiently reliable and robust. Thus, practitioners can use it to better plan the rehabilitation process of water mains and academics to get information for further studies.

Tran et al. (2007) compared Neural Networks Model calibration using Markov Chain Monte Carlo Simulation (Bayesian Weight Estimation) to Neural Network Models calibration using traditional Back Propagation algorithms to tackle the uncertainties in the latter one. This study focused on concrete storm water pipes in the City of Greater Dandenong (CGD) in Victoria, Australia, to evaluate the serviceability condition of storm water collector networks. The result of the study was also compared to that of Multiple Discrimination Analysis (MDA). This research indicated better performance of using Neural Networks Model with Bayesian Weight Estimation than Back Propagation and MDA. The authors, however, were not satisfied with the final result. Thus, they suggested that having a robust dataset may improve the accuracy of the model. Moreover, age was insignificant, whereas pipe size was reported as the most influential factor affecting the serviceability condition of the storm water network.

Asnaashari and Shahrour (2007) compared the result of ANN model to the Poisson regression model. The 10-year dataset of Sanandaj City-Iran was used to implement the examination. The authors noted that according to the result of this study, both models could predict water main failures. However, Artificial Neural Networks indicated a better performance compared to that of Poisson regression. It was also mentioned that one advantage of the ANN model is that it does not require any specific assumptions. Whereas, in the Poisson model, it is assumed that there is a logarithmic relationship between a dependent variable and independent variables.

In another study, to predict the rate of failure of water mains, Achim et al. (2007) applied ANN model to data from a specific utility in Australia. Cast Iron pipes with a diameter of less than 300 mm failed between 1997 and 2002 were examined. In this study, it was realized that ANN outperforms other statistical models when the dataset is relatively large, and more importantly, is noisy. Several pipe characteristics were employed in this study, such as length, material, diameter, and installation year. Nevertheless, the authors recommended that utilizing time-dependent factors, such as corrosion and climatic factors (seasonality), could improve the accuracy and reliability of the Neural Networks.

Fahmy and Moselhi (2009) investigated the remaining useful life of Cast Iron pipes with the application of two various neural network methods: the multilayer perceptron (MLP) and the general regression neural network (GRNN). The authors also compared the result with multiple linear regression to evaluate the accuracy and performance of ANN. Datasets from sixteen different utilities in the USA and Canada were obtained. Fourteen attributes were contributed to the analysis. For instance, in the GRNN model, authors found corrosion depth, pipe age, and soil resistivity to have had a significant impact on the remaining useful life of Cast Iron pipes. The importance of these factors was determined by calculating each input variable's contribution to producing output variable using Neural Networks. Researchers suggested that

this model be applied to either an individual pipe or even a group of pipes. In addition, the authors recommended that employing soil-related factors, such as soil corrosivity, soil resistivity, and soil pH, would result in more reliability. Therefore, the model may be used across the globe more efficiently.

Jafar et al. (2010) applied the artificial neural network method to a dataset from a city located in the north of France. In this study, the rate of failure and optimal replacement year for an individual pipe were the primary outputs of the ANN. Furthermore, the authors used the cross-validation method to evaluate the accuracy of the model. After analyzing the results from this study, ANN was suggested to be used to employ the best strategies that may help decision-makers offer better solutions to manage water networks in terms of maintenance and rehabilitation.

Another study in 2013 indicated better performance of ANN compared to the Multiple Linear Regression model (Asnaashari et al., 2013). In this study, researchers applied ANN to the water network in Etobicoke, Ontario in Canada. Many factors such as length, diameter, soil type, and material were used in this study. ANN represented 94% accuracy as opposed to MLR with 75% accuracy. Authors believed that although MLR has relatively good accuracy, predictive models would require higher accuracy in the real world. In terms of protection, Cement Mortar Lining (CML) and Cathodic Protection (CP) are believed to increase the useful life of water networks.

One study investigated the time to failure by analyzing different features (Harvey et al. 2014). The authors, in this study, applied ANN to the historical database of Scarborough district located in the Great Toronto Area (GTA), Canada. As a result, an artificial Neural Network was found to be an efficient way to predict time to failure. Moreover, using Cement Mortar Lining and Cathodic Protection reduced the rate of annual water pipes failures, and seasonal change was investigated to be an important factor influencing the rate of failure in Scarborough.

In Kingston (Ontario, Canada), Artificial Neural Networks were utilized to predict the number of failures (Nishiyama and Filion, 2014). Several attributes such as length, age, diameter, and soil type were considered for this study. ANN represented relatively good accuracy in this research. The decrease in the rate of breaks in the Kingston network was related to replacing older pipes and having an appropriate performance of existing pipes within the network. In another study, Kutyłowska (2015) applied Neural Networks to a Polish network to predict failure frequency. Researchers found that ANN applied by the Quasi-Newton method gives a satisfactory convergence while used for water networks.

Giraldo-González and Rodríguez (2020) made a comparison between different machine learning and statistical models. Dataset from Bogota in Columbia was used for the analytical process. Linear Regression, Poisson Regression, and Evolutionary Polynomial Regression were compared to Artificial Neural Networks, Bayesian, Support Vector Machine, and Gradient-Boosted Tree. Statistical models indicated acceptable results. Neural Networks, however, among machine learning models, did not present the best performance. In this study, as previous studies mentioned (e.g., Kettler and Goulter, 1985), the negative relationship between diameter and

failure rate was proven. In addition, other factors such as construction practices, corrosion process, and environmental conditions are believed to be important contributing factors to pipe failures. The given table below summarizes studies that employed artificial neural networks for the deterioration process of water mains.

TABLE 2.4 – SUMMARY OF STUDIES THAT USED ARTIFICIAL NEURAL NETWORKS (ANN)

Authors	Input Variables	Target	Accuracy
Moselhi and Shehab-Eldeen (2000)	Diameter, Material, Installation Year, Failures Records, Visual Records	Failure Classification	$R^2 = 0.98$
Ahn et al. (2005)	Diameter, Material, Length, Install Year, Soil Type, Failure Type, Break Records, Soil Features, Precipitation, Temperature, Water Quality	Relationship Between Cause and Nature of Failure	-
Najafi and Kulandaivel (2005)	Diameter, Material, Length, Install Year, Pipe Depth, Slope	Condition Rate of Sewer Pipes	$R^2 = 0.52$ to 0.98
Al-Barqawi and Zayed (2006)	Diameter, Material, Length, Install Year, Soil Type, Failure Type, Break Records, Pipe Depth, Wall Thickness	Condition Rate of Water Mains	$R^2 = 0.931$
Tran et al. (2007)	Diameter, Material, Install Year, Pipe Depth, Location, Number of Trees, Slope	Serviceability Deterioration	-
Raed et al. (2007)	Diameter, Material, Length, Install Year, Soil Type, Break Records, Pressure, Soil Features, Failure Type, Wall Thickness	Rate of Failure	-
Asnaashari and Shahrour (2007)	Diameter, Material, Length, Failure Type, Break Records, Pressure, Pipe Depth, Average Daily Traffic	Rate of Failure	$R^2 = 0.78$
Achim et al. (2007)	Diameter, Material, Length, Install Year, Soil Type, Failure Type, Break Records, Location	Rate of Failure	$R^2 = 0.679$
Al-Barqawi and Zayed (2008)	Diameter, Material, Length, Install Year, Soil Type, Break Records, Pressure, Soil Features, Type of Land, Water Level, Average Daily Traffic, Service Type, Cathodic Protection, Hazen-William Coefficient	Predict performance and condition of water mains	$R^2 = 0.982$
Tabesh et al. (2009)	Diameter, Material, Length, Installation Year, Pressure, Pipe Depth	Rate of Failure	-
Nasser and Saleh (2009)	Diameter, Material, Length, Install Year, Soil Type, Pressure, Soil Features, Structural Details	Predict wire break of PCCP	$R^2 = 0.994$
Fahmy and Moselhi (2009)	Diameter, Material, Length, Soil Type, Pressure, Soil Features, Type of Land, Pipe Depth, Average Daily Traffic, Structural Detail, Corrosion Depth	Remaining useful life of Cast Iron	$R^2 = 0.96$
Tran et al. (2009)	Diameter, Material, Length, Install Year, Soil Type, Failure Type Pressure, Soil Features, Type of Land, Water Quality, Number of Trees, Slope	Deterioration Pattern	-
Jafar et al. (2010)	Diameter, Material, Length, Install Year, Pressure, Pipe Depth	Failure rate and optimal replacement year	$R^2 = 0.972$
Shi et al., (2018)	Soil Resistivity, pH, Sulfide, Soil Moisture, Wall Thickness	Condition assessment	$R^2 = 0.52$

Giraldo-González and Rodríguez, (2020)	Age, Diameter, Length, Failure Records, Operational Attributes	Probability of Failure	$R^2 = 0.95$
Almheiri et al. (2020)	Material, Length, Diameter	Time to Failure	$R^2 = 0.84$
Snider and McBean, (2018)	Installation Year, Length, Diameter, Soil Type, Lining Status, Total number of failures	Time to Subsequent Failures	$R^2 = 0.76$

2.4 Other Emerging Machine Learning Applications

2.4.1 Logistic Regression

Logistic Regression, also called logit, or logistic model, is one of the most powerful algorithms for machine learning purposes that has been used in several studies for predicting water main failure (Motiee and Ghasemnejad, 2019; Robles-Velasco et al., 2020; H. D. Tran et al., 2009; Vladeanu and Koo, 2015). This algorithm was introduced and utilized statistically before applying it to water main networks (Berkson, 1944; Cox, 1958). In statistical studies, logistic regression is typically utilized to generate the probability of belonging to different classes, either as a binary (0, 1) or categorical (Yes, No) outcome (Cox, 1958). For example, in the case of the present study, the binary outcome would be broken (1) and Non-Broken (0) pipes. The regression seeks to find relationships between different independent variables, e.g., Material, Diameter, and a categorical/binary dependent variable. The probability of the dependent event is estimated by fitting data to the logistic (sigmoid) curve (Park, 2013). Where there is a dichotomous output variable, binary logistic regression is generally used. However, in some cases, there may be more than two classes, and in order to tackle this challenge, multinomial logistic regression should be employed (Park, 2013). The popularity of logistic regression is due to the fact that it provides the outcome values between 0 and 1, which is an indicator of the probability of belonging to one class. Given are the main equations assumed in the logistic regression function (Kleinbaum and Klein, 2010). For acquiring the logistic model, z is written as the linear combination of input variables as a linear regression format (equation 8), followed by transforming this linear model to the logistic function (equation 10).

$$z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \tag{3}$$

$$f(z) = \frac{1}{1 + e^{-(z)}} \tag{4}$$

$$f(z) = \frac{1}{1 + e^{-(\alpha + \sum \beta_n X_n)}} \tag{5}$$

Where z is a linear combination of input variables X (Diameter, Material, etc.), β is the regression coefficient, and α is the regression intercept (constant term). The given figure demonstrates the primary format of the logistic curve (Kleinbaum and Klein, 2010). The prediction of belonging to each class depends on the value of *f(z)*. Should the output be more than 0.5, the instance would belong to class 1, which in this domain, 1 is an indicator of a broken pipe.

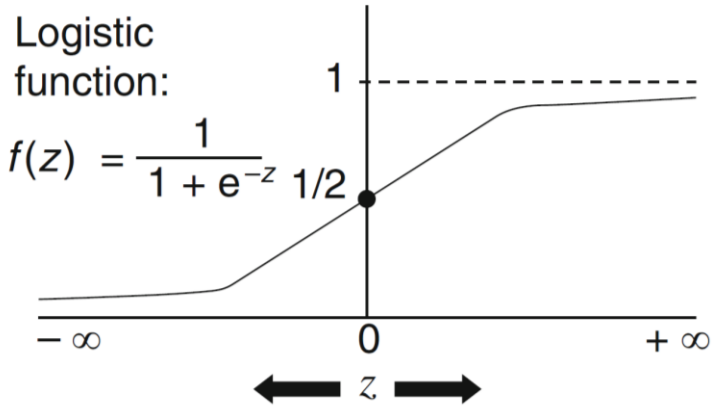


FIGURE 2.4 – THE PRIMARY FORMAT OF LOGISTIC CURVE (KLEINBAUM AND KLEIN, 2010)

The conditional probability can show the probability of belonging to class 1 (Broken Pipe) as below:

$$P(Y=1 / X_1, X_2, \ldots, X_n) = \frac{1}{1+ e^{-(\alpha + \sum \beta_n X_n)}} \quad (6)$$

$$P(X) = \frac{1}{1+ e^{-(\alpha + \sum \beta_n X_n)}} \quad (7)$$

2.4.2 Decision Trees

As the name suggests, Decision Trees (DT) has a tree-based structure and can be applied to both classification and regression problems (James et al., 2013; Swamynathan, 2019). For the pipe deterioration modeling, DT is used where the failure is imminent (Winkler et al., 2018). One of the most marked advantages of DT is its computational efficacy and also its straightforwardness (Breiman et al., 1984). The model forecasts target variables through a series of rules organized similarly to a tree (Syachrani et al., 2013). While training the model, building the rules starts at a first node known as the root, and it includes entire initially assigned observations (Swamynathan, 2019; Syachrani et al., 2013). Within the modeling process, each

branch is an indicator of the test outcome on the chosen attribute (True, False). Each leaf indicates the class of labels. The final decision is made after applying all calculations upon all attributes. The path which is produced from the root node to leaf nodes illustrates classification or regression rules. Therefore, a decision tree is made up of three main nodes: root node, branch node, and leaf node which represents the class of label (Swamynathan, 2019).

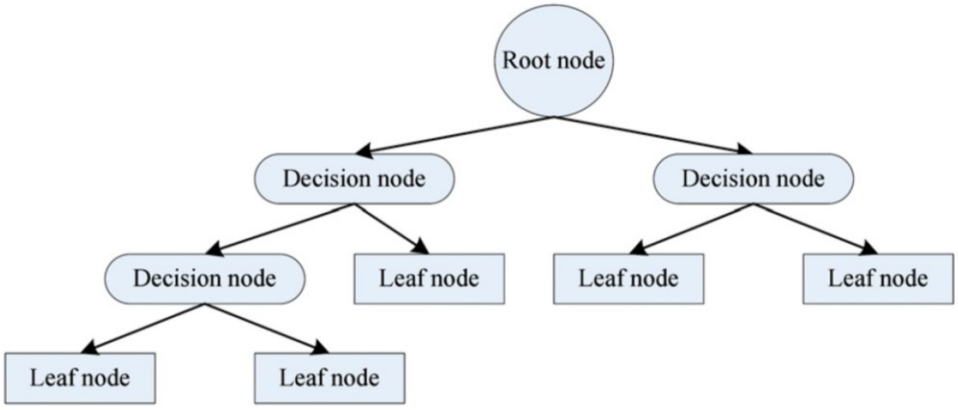


FIGURE 2.5 – CONCEPT OF DECISION TREES (SYACHRANI ET AL., 2013)

The output of a DT can be easily visualized and interpreted, helping practitioners perceive the most important factors that affect deterioration (Winkler et al., 2018). Swamynathan (2019) noted that a decision tree model utilizes the training dataset to build the tree model, and it ascertains which input variable should be employed for splitting the three into branches.

The greedy algorithm is the base of the decision tree method, upon which a tree is produced in a top-down recursive approach, and all training data points are located at the root node. Then, selecting an attribute, the tree model partitions the dataset into smaller portions. This splitting process is performed based on a statistical impurity measure known as Information (entropy) or Gini gain. Splitting stops when: a given node includes samples belonging to one class, no remaining attributes for continuing partitioning, and no more samples for partitioning. Then, given equations are employed to calculate the impurity within each node according to the selected attribute (Swamynathan, 2019).

$$Gini = 1 - \sum(p_i)^2 \quad (8)$$

Where pi is the probability of each class (in this domain, classes can be broken and non-broken water mains);

$$Entropy = - p \log2 (p) - q \log2 (q) \quad (9)$$

Where p and q show the probability of broken/non-broken pipes respectively in a selected node;

Several hyperparameters can be tuned to achieve an efficient DT model, such as maximum depth and maximum features, explained in further sections where necessary.

Having a significant variance is one of the main downsides of a simple decision tree (Hastie et al., 2009). In order to cope with this issue, bagging or bootstrap aggregation was introduced and can be employed to decrease the variance of the model (Breiman, 1996). This method can significantly improve the reliability and accuracy of the prediction. The training dataset, in this model, is portioned into multiple data points (samples) with replacement samples that have the same size as the original dataset (Swamynathan, 2019). Figure 2.6 shows the concept of bagging for a decision tree model.

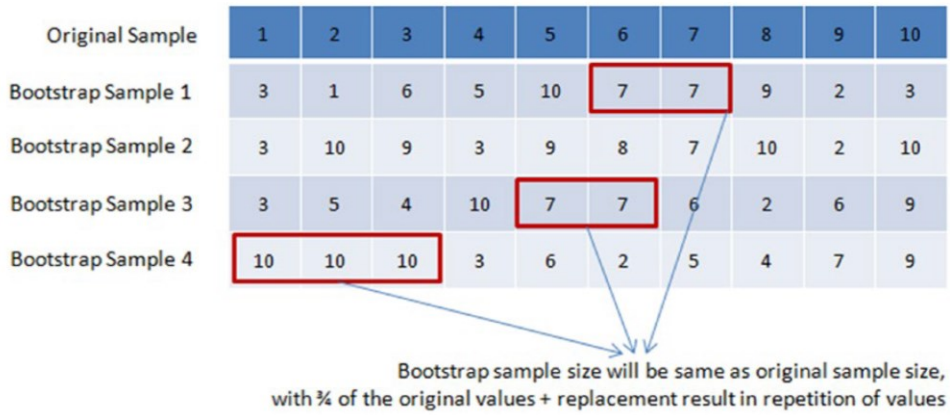


FIGURE 2.6 – BOOTSTRAPPING (SWAMYNATHAN, 2019)

In each step of bagging (bootstrap), independent models are created. It should be mentioned that for the regression model, the average of predictions, and for the classification model, the majority vote is considered for final prediction (Swamynathan, 2019).

2.4.3 Random Forests

Random Forests (RF) is another tree-based algorithm that is an ensemble machine learning method that can be used for both regression and classification analysis (CART) and is based on the combination of several decision trees (Breiman, 2001; Breiman et al., 1984). However, the conventional tree-based model inclines to overfit the training data set, leading to a weak performance (Murphy, 2012; Sadler et al., 2018). RF is an approach that could be utilized in order to address this issue. The given figure presents the concept of RF (Zhang et al., 2018)(Figure 2.7).

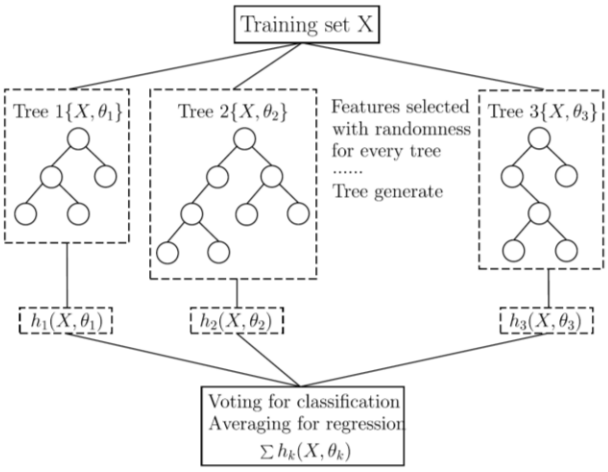


FIGURE 2.7 – RANDOM FOREST STRUCTURE (ZHANG ET AL., 2018)

In addition to bagging, which was introduced previously, RF improves the model's accuracy by applying minor tweaks, making trees decorrelated (James et al., 2013). Similar to bagging, several decision trees are built using bootstrapped (bagged) samples. However, when these trees are built (every time splitting is considered), m predictors are chosen as candidates from a set of p variables. The splitting process is allowed to employ only one of these m candidates, and this process continues freshly at each split (James et al., 2013). That is, at each split, the RF algorithm is only allowed to use the minority of predictors. This would help the algorithm avoid repeatedly using the most influential factors when creating decision trees and use all of the predictors during the modeling process. This is the main difference known between bagging and RF.

Random Forests, either classification or regression model, consider the average of all trees or majority votes, respectively, for final prediction (Breiman, 2001; Sadler et al., 2018). Selecting different variables randomly while creating decision trees, the RF can prevent overfitting by creating several weak learners that employ these random predictors.

Additionally, the RF can be used to extract feature importance from a dataset. Since many trees are created based on random predictors, RF can learn and record the significance of input features while performing prediction. Therefore, RF could be considered one of the most powerful algorithms for detecting the most critical factors (Sadler et al., 2018; Shirzad and Safari, 2019). Furthermore, the number of trees can be defined prior to the modeling process; the default value is 100 trees based on Scikit learn documentation.

2.4.4 GRADIENT BOOSTING

Boosting is another common approach that can be employed for improving the accuracy of any applied algorithms and is an intuitively similar approach to bagging (Freund and Schapire, 1999;

Hastie et al., 2009; James et al., 2013; Winkler et al., 2018). Freund and Schapire (1999) first introduced this method in their first boosting paper, introducing the renowned Adaptive Boosting algorithm known as AdaBoost (Freund and Schapire, 1997). This approach, theoretically, can be utilized to dramatically decline the error of any learning algorithm that produces models with a little more reliable and accurate performance than only random guessing (Freund and Schapire, 1996). Furthermore, this method can develop a robust predictive model from a list of weaker models by training each model (classification or regression) on a slightly changed subset of the dataset (Freund and Schapire, 1997). Essentially, this method transfers weak learners to significantly stronger learners that can be employed more reliably in the real world (Swamynathan, 2019). To clarify the concept behind boosting, let's assume $G_m(x)$ as sequential classifiers and take a_m as each classifier's weight created sequentially. The final robust model is produced with the combination of all models to a weighted majority votes (Winkler et al., 2018) (equation 13).

$$G(x) = \text{sign}\left(\sum_{m=1}^{M} a\, m\, Gm\,(x)\right) \tag{10}$$

Where G(X) is the final model; m is the number of an individual model, a_m is the weight of each model; and $G_m(x)$ is an individual model;

It should be noted, the main difference between boosting and bagging is the fact that the training dataset is resampled strategically to represent the most valuable information for each successive model (Zhang and Haghani, 2015). In boosting, misclassified samples have a higher probability of being chosen with a higher weight for making sequential models. Consequently, every freshly created model emphasizes those instances that have been classified incorrectly. The given figure indicates the process of boosting by one of the most powerful boosting algorithms called AdaBoost (Hastie et al., 2009) (Figure 2.8):

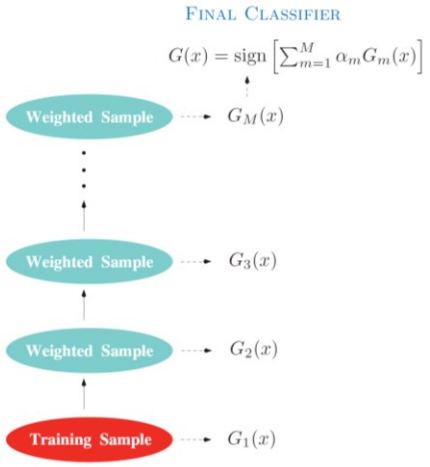


FIGURE 2.8 – ADABOOST FLOWCHART (HASTIE ET AL., 2009)

Gradient boosting (GB) is a state-of-the-art machine learning approach used widely for noisy datasets and datasets with an intricate relationship between attributes (Dorogush et al., 2018). This ensemble learning algorithm creates a robust predictive model combining several sequential weak learners (Friedman, 2001, 2002). Typically, this method is used for decision trees (Dorogush et al., 2018; Snider and McBean, 2018). However, GB can be employed for any machine learning algorithm to improve the performance and accuracy of final models. According to previous studies, the main difference between Adaboost and gradient boosting is weight adjustment. In Adaboost, weights are adjusted to minimize the number of misclassifications in classification problems or errors in regression problems. This process is executed based on the prediction of target values.

On the contrary, in gradient boosting, this reduction in misclassifications and errors is achieved based on the prediction of misclassified samples or residuals. Thus, gradient boosting tries to minimize wrongly classified samples or decrease the residuals in each iterative step of the process (James et al., 2013; Rebala et al., 2019; Swamynathan, 2019). In addition, gradient boosting utilizes the first or second derivative function format to better decline the misclassified samples.

2.4.5 Extreme Gradient Boosting Trees

Extreme Gradient Boosting Trees (XGBOOST) was first time introduced by Chen and Guestrin (2016). Like AdaBoost and GB, XGBOOST is an ensemble machine learning method trained and created by combining several decision trees (Chen and Guestrin, 2016). XGBOOST is the developed and more regularized version of the gradient boosting method (Swamynathan, 2019). This method, specifically, employs gradient descent to build sequential decision trees that decline residuals (Snider and McBean, 2018). This algorithm is reported to have less susceptibility to noises and outliers and a shorter time for the training process (Snider and McBean, 2020a). In one study conducted in 2018, XGBOOST was reported to have outperformed random forest and artificial neural networks for water main failure prediction (Snider and McBean, 2018). XGBOOST is known as one of the most efficient, largescale, and scalable prediction models (machine learning models) that have been contributing to winning solutions (17 out of 29 solutions) in Kaggle, where analytics competition in the data science field is performed (Swamynathan, 2019; D. Zhang et al., 2018). As can be seen at each iteration, the residuals will be employed to adjust the predecessor predictor to optimize the particular loss function (Zhang et al., 2018)(Figure 2.9).

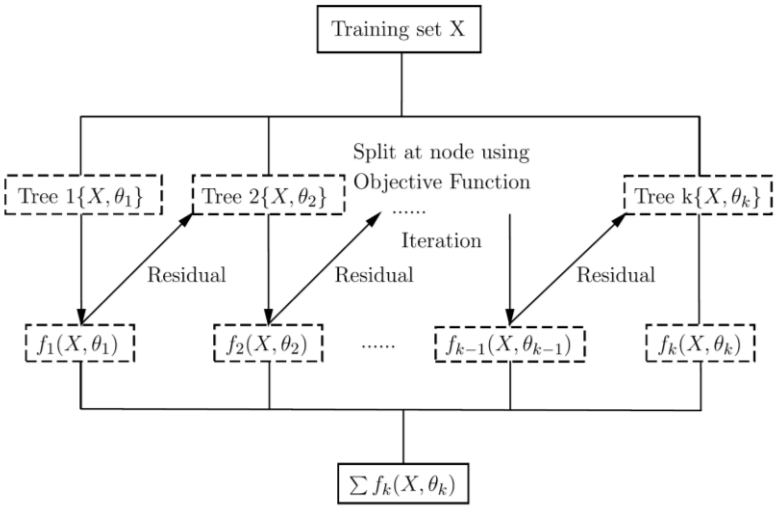


FIGURE 2.9 – FLOW CHART OF EXTREME GRADIENT BOOSTING

In order to achieve objective function, a regularization parameter is added to the loss function as an improvement for XGBOOST. The given equation indicates the objective function:

$$J(\Theta) = L(\Theta) + \Omega(\Theta) \quad (11)$$

Where Θ are parameters trained from data;

L is the training loss function (square loss, logistic loss) and an indicator to know how well XGBOOST fits on the training dataset;

After a specific number of node splitting in the decision tree, the improvement of model accuracy can be evaluated based on the objective function mentioned before. If there is an improvement in the model's performance after splitting the node, the change would be taken into account. Otherwise, the splitting process would not proceed. Provided information about XGBOOST is adopted from the main article (Chen and Guestrin, 2016) and one study conducted about Wind Turbines (Zhang et al., 2018).

2.5 COMPARISON OF PREVIOUS MACHINE LEARNING STUDIES

In addition to different probabilistic, fuzzy, and neural network approaches, some studies have been conducted using different machine learning algorithms, such as Random Forest, Naïve Bayes, and Logistic Regression. Accordingly, some of the most important related literature is provided in this section, and related results are presented briefly.

Almheiri et al. (2020) compared different machine learning models for time to failure prediction in Quebec City, Canada. In this study Ridge Regression (ℓ2), Artificial Neural Networks (ANN),

and Ensemble Decision Trees (EDT). Material, length, and diameter were used as input variables. The authors, in this study, utilized a Global Sensitivity Analysis (GSA) to evaluate the robustness of the models and indicate the correlation among different inputs and outputs. With considering all of these analytical processes, Cast Iron (CI), Hyperscon/Concrete (Hy), and Ductile Iron (DI) were found to be the most vulnerable type of materials for the output of models. They also found that material and length are the most contributing factors to water main failures. Therefore, in this study, EDT was recommended as a model that can predict the failure of water mains more accurately due to its efficient computational processes.

In another study in 2020, different machine learning classification models such as Artificial Neural Networks (ANN), Naïve Bayes (NB), Gradient Boosting algorithm (GB), and Support Vector Machines (SVM) were compared (Giraldo-González and Rodríguez, 2020). GB indicated the best performance among these classification algorithms with higher accuracy. This method also showed a better performance considering imbalanced datasets, which is a significant challenge for classification models. Furthermore, the models in this study can also predict the probability of failures for an individual pipe within the system.

Robles-Velasco et al. (2020) compared Logistic Regression (LR) and Support Vector Machines (SVM) as classification models to predict the probability of failures in one Spanish city. LR showed slightly higher accuracy and performance compared to SVM. From LR feature importance, it was referred that material is the most influential factor within the analyzed network, followed by many historical failures and length of segments. Another finding in this study was the fact that smaller pipes are more prone to failure.

A further study in 2020 compared one of the most powerful statistical approaches, Weibull Proportional Hazard survival analysis (WPH), to one of the most efficient machine learning algorithms called Extreme Gradient Boosting Decision Trees, known as XGBOOST (Snider and McBean, 2020). The main aim of this study was to forecast the time to subsequent failures in water mains. The results from this study recommended that machine learning approaches are more suitable for short-term planning, where the physical security of the network is of significant importance for the next several years. However, powerful statistical models such as WPH could be more favorable for long-term planning since corporates censored information to the analytical processes. Therefore, these authors investigated the feasibility of XGBOOST in a different study separately (Snider and McBean, 2018).

Random Forest (RF) and Multivariate Adaptive Regression Splines (MARS) were compared by Shirzad and Safari in 2019. A variety of features such as diameter, length, pipe depth, age, and average hydraulic pressure were considered for this study. The random forest model indicated better efficiency than the MARS model, although the results were close. In this study, datasets from Mahabad City and Mashhad City were used, and diameter, age, and hydraulic pressure were found to be the most important attributes within these networks.

Different tree-based algorithms were compared in another study by Winkler et al. (2018). The authors applied Decision Tree, Random Forest, Random under Sampling Boosting, and Adaptive

Boosting methods in order to predict the probability of failures. The authors recommended that tree-based algorithms could be an appropriate alternative to traditional statistical models. In this study, Boosted Random under-sampling method outperformed other methods.

Furthermore, one study in 2018 compared different machine learning models, such as Multiple Linear Regression, Artificial Neural Networks, Random Forest, Support vector machines, and Stacking Ensemble Regression. The main aim of this study was to predict the condition of pipes, considering different soil characteristics, including resistivity, pH value, Sulfide, soil moisture, and wall thickness. As a result, stacking regression was found to be the most efficient method to predict the condition of water pipes with a lower rate of errors (Shi et al., 2018).

Given table compares the abovementioned studies (TABLE 2.5):

TABLE 2.5 – COMPARING RECENT STUDIES CONDUCTED USING MACHINE LEARNING METHODS

Authors	Models	Input Variables	Target
Winkler et al., 2018	Decision Tree, Random Forest, Random Under Sampling, Adaptive Boosting (Adaboost)	Age, Material, Diameter, Pressure, Length, Service Type, Number of hydrants, …	Probability of Failure
Shi et al., 2018	Multiple Linear Regression, ANN, Random Forest, SVM, Stack Regression	Soil Resistivity, pH, Sulfide, Soil Moisture, Wall Thickness	Condition assessment
Shirzad and Safari, 2019	Multivariate Adaptive Regression, Random Forest	Diameter, Length, Pipe Depth, Age, Hydraulic Pressure	Rate of Failures
Almheiri et al., 2020	Ridge Regression, ANN, Ensemble Decision Trees	Material, Length, Diameter	Time to Failure
Giraldo-González and Rodríguez, 2020	ANN, Naïve Bayes, Gradient Boosting, SVM	Age, Diameter, Length, Failure Records, Operational Attributes	Probability of Failure
Robles-Velasco et al., 2020	Logistic Regression, SVM	Material, Diameter, Age, Length, Connections, Pressure Fluctuation, Total number of failures	Probability of Failure
Snider and McBean, (2018)	ANN, XGBOOST, Random Forest	Installation Year, Length, Diameter, Soil Type, Lining Status, Total number of failures	Time to Subsequent Failures

In summary, the abovementioned studies tried to employ different factors affecting the deterioration process of water pipes. However, the number of case studies was limited, and the authors applied the model to a few utilities. Therefore, it is not justifiable to utilize one global approach for all utilities, and each network must be investigated separately. Additionally, most of these studies did not consider censored information related to replaced pipes or the pipes

installed before collecting information. Accordingly, Snider and McBean (2020) recommended that these state-of-the-art models be integrated with more advanced statistical models such as survival analysis to have a better long-term plan for maintaining and keeping water networks in a desirable condition. Finally, many of these studies did not integrate the results into GIS files and did not consider some of the most critical factors, such as hydraulic pressure or joint types.

2.6 FACTORS AFFECTING PIPES DETERIORATION

A good understanding of the different factors that affect pipe failure is key to analyzing pipe deterioration. This can also help data scientists create more reliable predictive models and thereby decrease the cost of maintenance and replacement of water mains (Barton et al., 2019).

Shamir and Howard classified pipe deterioration factors into four main categories:
1. Age of pipes and their manufacturing quality, including pipe components such as connectors and material.
2. Environmental factors, such as frost load, traffic load, and pipe-soil interaction;
3. Installation practices;
4. Operational conditions, such as water pressure as well as water hammer.

Kleiner and Rajani classified deterioration factors into three categories: static, dynamic, and operational. Static factors are those related to physical characteristics which do not change significantly over time, such as diameter, material, wall thickness, installation practices, and soil type. On the other hand, dynamic factors change over time, such as temperature, soil moisture, soil electrical resistivity, bedding condition, age, and dynamic loadings. Additionally, cathodic protection, replacement, previous failures, and water pressure are operational factors. This aligns with the InfraGuide (2003) definition of operational factors, including internal and transient pressure, water quality, leakage, flow velocity, backflow potential, and O&M practices. However, some of the factors defined as static by Kleiner and Rajani can change over time, with deterioration and material build-ups, such as diameter and wall thickness.

Based on a review of previous studies, Barton et al. (2019) categorized the most important factors into three groups: intrinsic, operational, and environmental factors. The authors explored factors impacting the most commonly used pipe materials (Cast and Spun Iron, Ductile Iron, PVC, Polyethylene, Steel, and Asbestos Cement). This classification is similar to that of Shamir and Howard (1979) but groups age, manufacturing quality, installation practices, and other initial pipe characteristics into intrinsic factors. Given its simplicity and broad applicability, the current study will follow the Barton et al. (2019) classification.

Barton also analyzed the interaction and effectiveness of different factors on each other. For instance, a dramatic seasonal change could lead to decreasing or increasing soil moisture. This may lead to ground movement and cause premature failure. The following sections explore each factor and its impact on main breaks based on a literature review.

2.6.1 Intrinsic Factors

2.6.1.1 Material

Various types of pipe materials have different probabilities of failure. For example, iron-based pipes (Cast Iron, Spun Iron, and Pit Cast Iron) are more prone to internal corrosion due to weaker structural integrity (Clair and Sinha, 2014). Cast Iron (CI) was introduced as a pipe material around the beginning of the 19th century when iron was molten and cast in sand molds (Clair and Sinha, 2014). However, the casting process could lead to a non-uniform thickness, which is more prone to failure. To mitigate the challenges associated with CI, spun gray iron was released during the 1930s. Later, Ductile Iron (DI) and Steel pipes were introduced as alternatives which were typically more resistant to failure. The main difference between CI and DI is the shape of graphite, which is spherical in DI, and flakes in CI pipes (Barton et al., 2019). This makes DI structurally tougher compared to spun and cast iron. Steel pipes, however, are generally stronger than DI but less resistant to corrosion. Therefore, DI is typically preferred for pipe diameter between 300 to 800 mm, and ST pipes for over 800 mm. Concrete-based pipes such as Asbestos Cement (AC) are more corrosion resistant but less flexible and may fail more easily with underground movement than DI and steel pipes (Mordak and Wheeler, 1988). In recent decades polyvinyl chloride (PVC) has emerged as a popular pipe material and an alternative to AC pipe (Barton et al., 2019; Røstum, 2000). PVC is much more resistant to corrosion and more flexible than AC. PVC pipes also have generally lower manufacturing costs and more straightforward installation. Polyethylene (PE) pipes are another plastic pipe option that can withstand higher pressures than PVC and are thus more durable (Barton et al., 2019).

Folkman (2012) reported that approximately 80% of pipes installed in the USA and Canada were a combination of CI (28%), DI (28%), and PVC pipes 23%. This statistic then was confirmed by similar research (Folkman, 2018), which indicated that just over 90% of current pipes are CI (28%), DI (28%), PVC (22%), and AC (13%), and the remainder of pipes utilized were denoted by HDPE, steel, molecularly oriented PVC (PVCO), concrete steel cylinder (CSC), and other materials.

2.6.1.2 pipe age

Watermain rate of failure was initially conceptualized to have an exponential or linear relationship with age (Shamir and Howard, 1979). However, other studies have proven that the impact of age on failure is complex (Andreou, 1986; Andreou et al., 1987a; Jeffrey, 1985). For example, Andreou et al. (1987) reported that failure rates differ by pipe vintage due to changes in manufacturing, standards, and construction practices. For instance, new casting methods led to grey cast iron pipes with thinner walls which became more prone to failure when exposed to corrosive soil and external loads (Røstum, 2000).

In some cases, the failure rate is high at the beginning of the bathtub curve mentioned before due to improper installation, improper bedding, or pipe deficiency. In one study carried out by Folkman in 2018 on data from various utilities in Canada and the USA, the average failure age

was reported to be around 50 years. Barton et al. (2019) noted that the failure rate has steady progress in some types of pipes, e.g., PE and PVC due to their different structures. Plastic-based pipes are not prone to electrochemical corrosion. Therefore, there is no specific mechanism for these pipes to experience structure deterioration over time (Task Committee on Water Pipeline Condition Assessment, 2017). Thus, age can be used as an indicator of the continued impact of environmental variables, when more information is not available. However, accounting for the conditions that lead to failure at different ages should produce better results.

2.6.1.3 PIPE DIAMETER

The pipe break rate has been found to be higher in smaller pipes than larger pipes (Kettler and Goulter, 1985). Larger pipes generally have thicker walls and more reliable joints, making them more resistant to ground movement and corrosion (Wengström, 1993). Their failure will also undoubtedly lead to more significant consequences than smaller pipes, which also means they are more likely to receive preventative maintenance and rehabilitation. Furthermore, smaller pipes are more common in water distribution systems, and more data is available on them. Folkman found that almost 67% of pipes across the US and Canada diameter less than 200 mm. Hence, it is important to include smaller pipes in failure analyses and consider diameter as a factor.

2.6.1.4 JOINT SYSTEMS

Joint failure is one of the frequent types of failures for water distribution networks. A few studies reported that 15% and 16% of PVC pipes on average fail as a result of joint failure (Dingus et al. 2002; Burn et al. 2005). In addition, Folkman (2018) highlighted that joints could be one of the main sources of leakage. Fittings and joints can be an integral part of a network, built as a pipe, or non-integral, connecting the ends of two pipes. Joints can also be categorized as rigid or flexible. Rigid joints are less flexible and more likely to fail due to dynamic conditions, such as ground movement, traffic load, and water pressure.

Moreover, this type of joint is susceptible to failure due to poor installation and corrosion at the connections (Ruchti, 2017). Flexible joints, on the other hand, allow the pipe to move marginally when confronting dynamic conditions. Accordingly, flexible joints are more desirable in areas where ground movement is likely (Barton et al., 2019; Ruchti, 2017). Nevertheless, poor installation and sudden ground movement may still lead this type of joint to fail.

2.6.1.5 PIPE COATING AND LINING

Coating (outer surface) and lining (inner surface) are methods typically used to protect pipes against corrosion (InfraGuide, 2003; Mordak and Wheeler, 1988). These methods were first used to protect CI pipes and later DI and steel (Barton et al., 2019). AC pipes are only recommended to be coated where soils surround them with a pH below 6.0, although concrete

is generally naturally resistant to corrosion (Trew et al., 1995). Various types of lining and coating can be applied to different materials, such as bitumen, cement mortar, synthetic resin, PE sleeve, Epoxy, cathodic protection, and coal tar. AC pipe, for instance, is generally coated in bitumen or coal tar. (Farrow et al., 2017). It should be mentioned that the deterioration process of pipes could be affected by different types of protections; thus, the provided information is of significant importance for infrastructure analysts (Barton et al., 2019).

2.6.1.6 MANUFACTURING DEFECTS

A wide range of defects may increase the likelihood of pipe failure (Barton et al., 2019), such as non-uniform wall thickness, porosity, inclusions, micro-cracks, and so forth. For instance, porosity emerges when air is trapped in the mold while molten iron is solidifying. Similar defects can also occur in PVC pipes. These imperfections may cause a weakened point, micro-cracks, or even a fracture. Additionally, in some types of pipes, these deficiencies result in strength reduction. The problem may be addressed by considering stricter quality control criteria and enhancing the quality of manufacturing. In order to mitigate these defects, more stringent quality control may be taken into consideration.

2.6.1.7 PIPE DAMAGE FROM HANDLING, STORAGE, AND THIRD PARTIES

After manufacturing, pipes may still be damaged in storage, transportation, installation, and operation. As Barton et al. mentioned, the coating of pipes can be damaged due to poor handling. This means cracks and dents could emerge on the coating layer. It was also noted that some types of pipes, for example, DI, may easily dent under excessive force due to the wall's thinness. AC pipes and other various metal pipes are also likely to be damaged in poor handling conditions.

Moreover, most plastic pipes are susceptible to embrittlement while exposed to ultraviolet light for an extended period (Barton et al., 2019). Hence, the storage process of these pipes is vital. Third-party activities around the pipes can also affect their structure (Røstum, 2000). For example, excavation in the vicinity of a pipeline may alter the bedding shape, leading to differential settlement and eventually failure.

2.6.1.8 LENGTH

The effect of this attribute has been investigated in many studies before (Andreou et al., 1987b; Andreou, 1986; Aydogdu and Firat, 2015; Kleiner and Rajani, 2001, 2002; Philip and Aljassmi, 2020; Rajani and Kleiner, 2001; Rajeev et al., n.d.). Not only to find the impact of this factor but also for data engineering purposes and creating rate of failures. Philip and Aljassmi (2020) found length as an attribute that should be taken into account while trying to make a predictive model. Aydogdu and Firat (2015) noted that the failure rate was higher for pipes with lengths of 0 to 200 m. Therefore, the present study utilized length for creating the current rate of failure. This attribute has also been used as an input variable for creating classification models.

2.6.2 Environmental Factors

2.6.2.1 Soil Moisture, Ground Movement and High Temperature

Soil moisture is an environmental factor that may cause different types of failures, such as corrosion and ground movement. One study found the pipe rate of failure to be highest during mid to end of summer (Gould et al., 2011). During this period, soil moisture is reduced, leading to soil shrinkage and ground movement and shrinkage. This movement would lead to a higher rate of circumferential breaks. In another study, the soil moisture deficit (SMD) contributed to the higher rate of failure, especially during summer for Asbestos Cement pipes (Pritchard et al., 2013). Wols and Thienen (2014) reported a similar result considering temperature and lack of rainfall. The failure rate for AC and steel pipes was higher in the summer than in other seasons. Barton et al.(2019) found that the pipe rate of failure was lowest during the spring season when soil is typically wet, and ground movement is less likely to happen. Similarly, it is generally highest at the end of summer and beginning of fall, when SMD is highest (Chan et al., 2005).

In addition to shrinkage, other soil-related movements include sand washout and compaction (Pritchard and Hallett, 2013). One study found that the frequency of failures had a temporal-spatial relationship with the failures that happened previously, especially in sandy soils. Sand wash-out is a common challenge that may lead to pipe failure in a long time. Moreover, the natural process of soil compaction and construction of poor bedding and foundation would typically lead to differential settlement, leading to premature pipe failures (Wols et al., 2014). With the effects of climate change and increased drought periods, pipe differential settlement is recommended that climate changes and the likely surge of the drought period will probably increase the effect that differential settlement has on pipes failures (Wols and Thienen, 2014a).

2.6.2.2 Other corrosion-related Soil Characteristics (pH, Resistivity, Corrosivity, etc.)

Corrosion is one of the main causes of pipe deterioration (Barton et al., 2019). Folkman reported that corrosion accounts for approximately 28% of pipe failures. Wasim et al. also noted that soil corrosivity is one of the most contributing factors to metal pipe deterioration. It has also been the primary cause of the deterioration of CI water mains (Wang et al., 2016). The deterioration of metallic pipes, especially CI, can be observed external or internal corrosion (Rajani and Kleiner, 2013). External corrosion happens due to soil-pipe interactions, whereas internal corrosion is a result of water-pipe interactions. Corrosion, in the end, may affect the material integrity of a pipe by weakening the wall thickness, which may lead to failure (Barton et al., 2019; Doyle et al., 2003). Soil characteristics, such as pH, electrical resistivity, moisture content, and sulfates, directly impact the grade and type of corrosion (Doyle et al., 2003). Depending on the location of the study, these soil characteristics may vary noticeably. The given figure indicates different factors which contribute to external corrosion of water mains

(Wasim et al., 2018) (Figure 2.10). Figure 2.11 also depicts the corrosiveness scoring of soil, published by AWWA (Doyle et al., 2003).

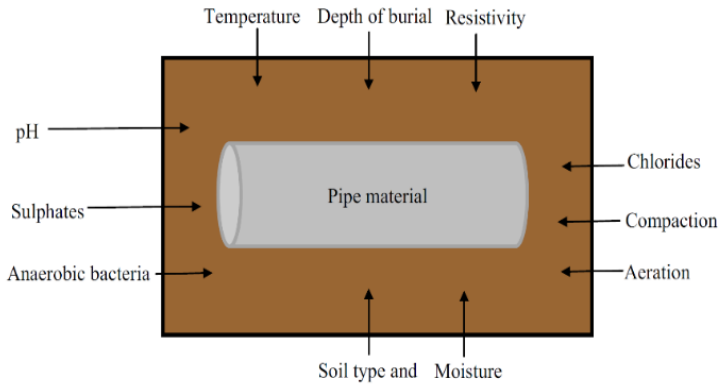


FIGURE 2.10 - FACTORS LEADING TO CORROSION OF WATER MAINS (WASIM ET AL., 2018)

Soil characteristics	Points*
Resistivity ** (Ω.cm)	
< 700	10
700 - 1000	8
1000 - 1200	5
1200 - 1500	2
1500 - 2000	1
> 2000	0
pH	
0 - 2	5
2 - 4	3
4 - 6.5	0
6.5 - 7.5	***0
7.5 - 8.5	0
> 8.5	3
Redox Potential	
> + 100 mV	0
+50 to +100 mV	3.5
0 to +50 mV	4
Negative	5
Sulphides	
Positive	3.5
Trace	2
Negative	0
Moisture	
Poor drainage, continuously wet	2
Fair drainage, generally moist	1
Good drainage, generally dry	0

* Ten points means that soil is corrosive to grey or ductile cast iron pipe: protection is indicated
** Based on single-probe at pipe depth or water-saturated soil box.
*** If sulphides are present and low or negative redox-potential results are ontained, give three points for this range.

FIGURE 2.11 - AWWA SOIL CORROSIVENESS SCORING SYSTEM (ANSI/AWWA C105/A21.5-99)

Corrosion is categorized into two main types: graphitization and corrosion pitting. Graphitization usually happens in CI pipes, and it happens when graphite in iron alloy reacts with iron oxide (Barton et al., 2019). On the other hand, corrosion pitting happens in different types of steel and iron pipes. Although graphitization cannot be readily detected, it strongly affects the structure of the pipe, leading to mechanical failure (Ruchti, 2017). The corrosion of CI Pipes leads to thickness decrease, pitting, and the emergence of graphitic areas. This graphitic area, in some cases, makes visual inspection difficult. Therefore, the inspection could be done when the layer is removed (Wang et al., 2018).

The pH value is also considered to be another factor that affects corrosion of CI water mains either indirectly or directly (Wang et al., 2018). Other studies indicated that when the soil pH decreases, the rate of corrosion increases (Petersen and Melchers 2012; Kreysa and Schütze 2008; Nesicet al. 1996). It was also mentioned that the corrosion rate is not related to pH when soil pH is over 5. Generally, soil with a lower pH rate is considered more corrosive than soil with

a higher degree of pH. Nevertheless, soil with a pH of 5.5 to 8.5 may cause extreme corrosion in some specific environments (Doyle et al., 2003).

Since corrosion is an electrochemical reaction, soil resistivity is known to play an important role in the corrosion of underground pipes (Doyle et al., 2003). Soil resistivity measures how resistant a specific type of soil is to the flow of electricity. McMullen (1982) suggested a linear relationship between pipe failure and soil characteristics and proposed a model applied to the water distribution network of Des Moines, Iowa. The researchers found that 94% of failures were attributed to soil with less than 2000 Ω-cm saturated resistivity, confirming corrosion as a major driver of deterioration. Another study found that pipe expected service life reduced by 28 years for every 1000 Ω cm decrease in soil resistivity (Kleiner and Rajani, 2001). However, McMullen (1982) noted that soil resistivity might not be constant since it is affected by many factors, such as road salting, acid rain, and temperature. (Bhattarai, 2013)(TABLE 2.6).

TABLE 2.6 – THE RATE OF SOIL CORROSIVITY BASED ON SOIL RESISTIVITY (BHATTARAI, 2013)

Soil Resistivity (Ω.cm)	Soil Corrosivity Rate
> 20,000	Essentially non-corrosive
10,000 – 20,000	Mildly Corrosive
5,000 – 10,000	Moderately Corrosive
3,000 – 5,000	Corrosive
1,000 – 3,000	Highly Corrosive
<1,000	Extremely Corrosive

In addition to the factor mentioned above, some other factors such as stray current, the pressure of the atmosphere, and temperature could influence the process of corrosion in Cast Iron pipes. For instance, the temperature is believed to expedite corrosion reaction (Doyle et al., 2003). In addition, Gummow and Eng reported that the stray current became a major factor that led to electrolysis in buried water pipes due to the advent of electrical transit networks in North America. Hence, these factors should also be taken into consideration when analyzing corrosion in Cast Iron pipes.

2.6.2.3 SEASONALITY

The effect of unexpected and drastic seasonal changes has been investigated in previous studies and reported strongly correlating with the rate of failure. Failure rates have been reported to be higher during summer and fall, in which the weather is dry, especially for PVC and AC pipes, or during freezing winters for Iron and Steel pipes. The inverse condition is also reported for wet and mild weathers for fall and winter, respectively (Fuchs-Hanusch et al., 2013; Gould et al., 2011; Laucelli et al., 2014; Pritchard and Hallett, 2013; Wols and van Thienen, 2014). The failure rate is also reported to be higher in rigid mains up to 200 mm throughout the winter season, and one study found that the circular crack is more prevalent in pipes with the size of 150 mm during the summer season (Andreou et al., 1987a, 1987b; Fuchs-Hanusch et al., 2013). One specific study showed that seasonal changes do not significantly

impact pipes at depths of greater than 1 meter, particularly for larger pipes (Wengström, 1993). This is because larger pipes generally fail under pressure and not the movement of the ground, which may happen due to seasonal alteration. Wols and Thienen (2014) reported the most important weather features that may affect the deterioration process of water mains as temperature, frost, and deficit which typically happens as a result of seasonal changes; however, the consequences of these factors may vary geographically, and it usually depends on the type of materials (Gould et al., 2011; Laucelli et al., 2014).

2.6.2.4 COLD TEMPERATURE

The temperature may affect the deterioration process of water mains. Accordingly, many studies analyzed the failures caused by temperature changes in different seasons (Fuchs-Hanusch et al., 2013; Gould et al., 2011; Hu and Hubble, 2007; Rajani and Kleiner, 2001), which indicates that approximately 60% of failures would happen during the winter season (Rezaei et al., 2015). One study in the Netherlands found that different materials have different reactions to temperature, and Iron is reported to be the most vulnerable material for cold temperature (Wols et al., 2019). During cold winter, soil moisture may cause frost heave (Barton et al., 2019). This frost is reported to have a significant impact on pipes laid down shallower than 0.5 to 1 meter, depending on the duration of the frost (Pritchard and Hallett, 2013). In addition, bedding type may affect the impact of frost heave, leading to a higher rate of failure in water mains (Bruaset and Sægrov, 2018). These studies indicate the importance of temperature and its effect on the deterioration of water pipes. Therefore, it is worth including this feature in the prediction of future failures.

2.6.3 OPERATIONAL FACTORS

2.6.3.1 INTERNAL WATER PRESSURE

One study in 2019 reported the effectiveness of changes in water pressure which may increase the probability of failure (Barton et al., 2019). Changing operational pressure from consumer usage to the alteration in the network by management leads to a fluctuation in the current pressure, leading to fatigue failures (Rezaei et al., 2015). One conducted study in Zagreb represented the impact of pressure regulation on decreasing pipe failures by around 17%, which was clear for PVC and Iron pipes (Iličić, 2009). Water hammer (Transient surge pressure) also occurs due to operational events such as testing fire hydrant or pumping station failure. This additional pressure causes sudden changes in operational pressure inside the network, resulting in failures (Barton et al., 2019; Martínez-Codina et al., 2016). Martínez-Codina et al. in 2016 listed a range of materials such as metal and cement to be considered while taking pressure into account and reported that tolerance of pressure might lead to a higher rate of failures in larger diameter pipes, particularly in the places that resistance of pipes has weakened as a result of corrosion and micro cracks. One study investigated the fact that the smaller pipes are less prone to failures due to the alteration of operational pressure since they

typically do not experience a high rate of transient surge pressure (Ruchti, 2017). The studies mentioned above have emphasized the significance of operational pressure within the network. Accordingly, this factor should be considered important, especially for transmission networks where the water pressure may cause drastic internal pressure.

2.6.3.2 PREVIOUS FAILURES

The importance of previous failures has been investigated in previous studies. Clark et al. (1982) reported that initial breaks might result in subsequent failures that are located, spatiotemporally, close to the previous failures. Previous studies have proven that 22% of failures happened within the proximity (1 m) of previous incidents, and also 42% of these failures happened in the one-day interval after initial failure (Goulter et al., 1993; Goulter and Kazemi, 1988). Therefore, considering historical failures in water main failure prediction models is critical.

3. METHODOLOGY

This study applied various steps to predict water main deterioration, ranging from defining objectives, data collection, cleaning, and preparation, applying different models, evaluating and comparing models. These steps are shown in Figure 3.1 and described in more detail in the following sections. The methodology follows the Cross-Industry Standard Process for Data Mining (CRISP-DM). This process was introduced by ESPRIT (European Strategic Program on Research in Information Technology) (Swamynathan, 2019; Verdhan, 2020) as a global solution that does not consider any domain-dependent assumptions. The CRISP-DM includes business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

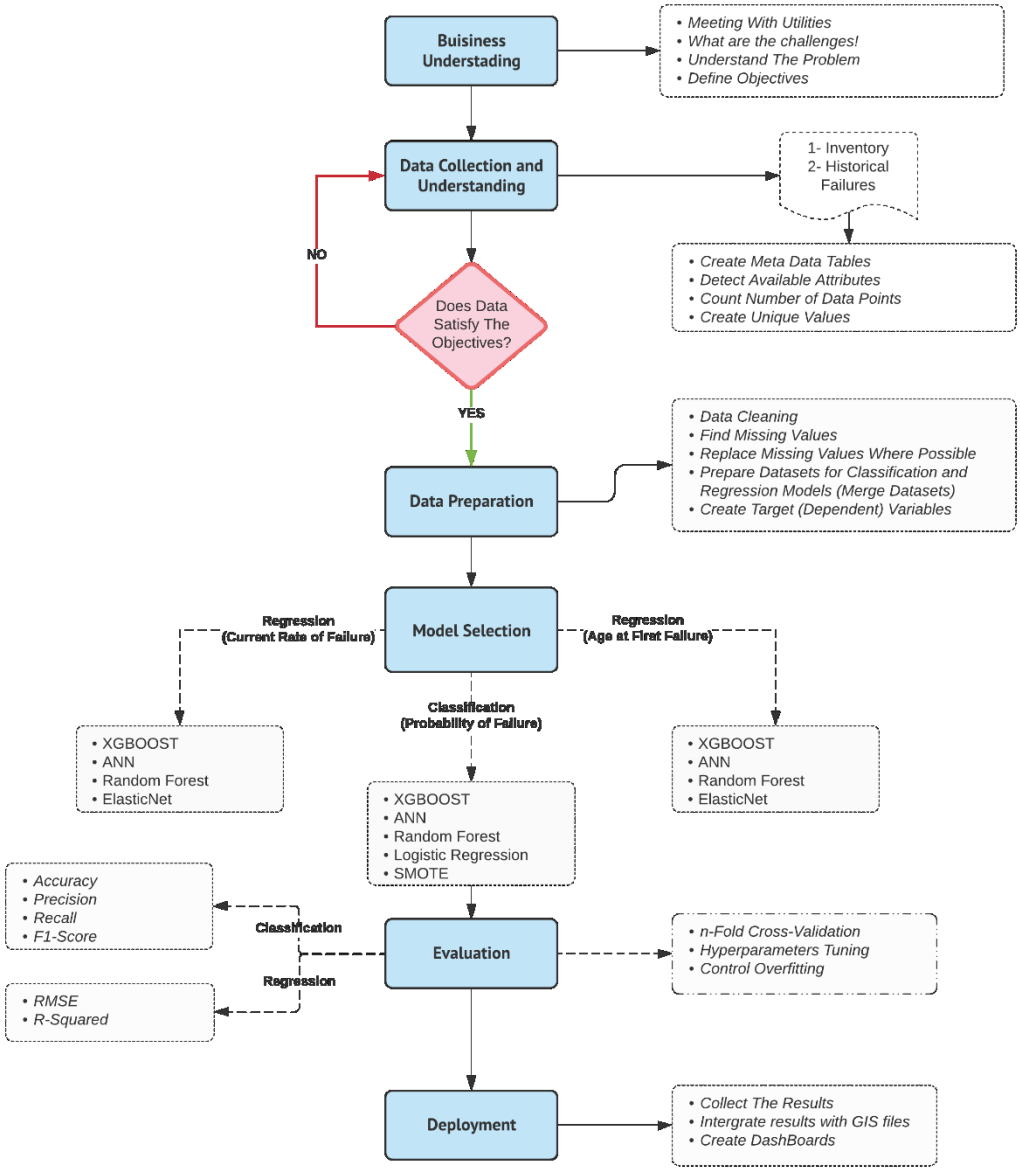


FIGURE 3.1 – FLOWCHART OF METHODOLOGICAL STEPS

3.1 Data Collection and Business Understanding

Each water utility provided two datasets, a water main inventory and a historical record of water main breaks. More details regarding the datasets and available attributes are provided in Appendix A. In order to better understand the available data as well as the priorities and challenges of the utilities, three workshops were organized with the National Water and Wastewater Benchmarking Initiative (NWWBI). During these meetings, three prediction targets were found to be preferred by the participating utilities. First and foremost, the probability of failure was chosen as the most important target since it can be directly used in asset management risk analyses. In order to predict this target, classification algorithms were selected. Some utilities were also interested in predicting age at first failure and rate of failure. The first failure is useful in estimating the expected service life of water mains since the likelihood of failure increases once a pipe has broken. The rate of failure can also be applied in estimating yearly maintenance and repair needs. These two targets were predicted with regression models. Details about the three targets and the models developed for each are provided in the following sections.

3.2 Data Understanding and Data Preparation

Data cleaning is one of the most important steps in the data mining process and can also be one of the most time-consuming. This step started with the development of metadata tables to describe the available raw data and ensure all information would be interpreted correctly in the following steps, from primary datasets. A wide range of attributes was provided for each dataset, which varied between different utilities since not all utilities collect various features related to the pipeline. However, some of these attributes, such as material and diameter, are common between inventory datasets and breaks datasets. It should be noted that the datasets employed in this study are collected from thirteen different municipalities across Canada. In order to reach a better understanding of all provided attributes, different metadata tables were created for different utilities. For each metadata table, some columns were defined in order to develop consistency between datasets. These columns are as follows, and for each column, a related description is provided (TABLE 3.1):

1. ***TITLE:*** This column includes the actual name of the attribute in the raw dataset, and it was created to compare the existence of different attributes between utilities and use the column as a reference for each dataset for further application.

2. ***Description:*** A brief explanation about the attribute to better perceive the value of the feature

3. ***Type:*** Type of attributes are provided in this column. For instance, categorical, polynomial, and numerical formats are defined for different features.

4. ***Name:*** This column is one of the most important ones, including the defined name for each attribute. This importance is because some attributes had different names within the dataset, although having the same values. In order to address this challenge, the "Name" column was created, which creates consistency between the datasets, and all required columns for the analysis at the end will have these names that are defined in the "Name" column.

5. ***Unit:*** This column includes units pertinent to different attributes

6. ***Range:*** The distribution of values for different attributes is included in this column

7. ***Category:*** If a specific attribute is a categorical type, different categories are listed in this column

TABLE 3.1 – SAMPLE OF METADATA TABLE

TITLE	Description	Type	Name	Unit	Range	Category
HISTKEY	Work Order Number	Numerical (Discrete)	WONumber	N.A	30099-1211812	N.A
ADDDTTM	Work Order Created Date	Numerical (Date)	WODate	Date	-	N.A
COMPKEY	Water Main Asset ID	Numerical (Discrete)	AssetID	N.A	1-938898	N.A
INITDTTM	Work Order Initiated Date	Numerical (Date)	BreakDate	Date	-	N.A
MODDTTM	Last Date Work Order	Numerical (Date)	WOmodification	Date	-	N.A
SCHEDDTTM	Scheduled Work Data	Numerical (Date)	ScheduledWork	Date	-	N.A
STARTDTTM	Work Started Date	Numerical (Date)	WOstart	Date	-	N.A

According to the provided metadata tables, inconsistencies and possible problems were detected from the data. A unique name is defined for each category, and unique IDs are identified for further matching break and inventory information. Any inconsistencies and outliers identified and removed from data.

The next step after creating the metadata table was creating cleaned data for all attributes within different datasets. This step has been performed in order to create groups of consistent values among all utilities. Almost all attributes were cleaned, and detailed categories were defined for different features. For instance, different categories with the same values were created in terms of material or cause and nature of failures, and there were significant inconsistencies within different datasets. Some utilities had these attributes as codes (which were not decoded at the beginning). This issue made the cleaning process more challenging. In addition, some causes of failures are not related to the natural deterioration or might not be related to the pipe itself. These elemental failures are not directly related to the pipe itself. Thus, some unique names were defined to distinguish these defects from pipe-related ones, which will be discussed in more detail. This cleaning process has led to an array of consistent datasets that can be efficiently utilized in data analysis.

The next step was applying these metadata tables and cleaned values to the main datasets. To do so, Power BI software has been used. All columns have been cleaned, and values are replaced with cleaned values that were created in the second step.

Additionally, some attributes are not available in all utilities. Therefore, these features are kept for each utility separately; however, while merging all cities, these attributes may be removed since their existence may affect the reliability and accuracy of the predictive models. Roughness and HGL, for instance, are among these attributes which are specific to some utilities. After cleaning the cause and nature of failures, this step has been the most time-consuming data preprocessing step. This is since all utilities should have been cleaned and made ready to start the preliminary analysis.

Supervised models that predict whether or not a pipe will fail or the probability of failure requires data on failed and non-failed pipes. Thus, the two available datasets, inventory, and breaks were merged for each utility. A unique ID was identified for each utility in order to match the two datasets. If a unique ID was unavailable or the matching percentage was low, GIS data was used to spatially join the datasets. The years of historical break data available for each city are listed in TABLE 3.2, as well as the matching percentages.

TABLE 3.2 – PERCENTAGE OF PIPE ID MATCHED BETWEEN INVENTORY AND HISTORICAL FAILURES

Utility	Years of Breaks	% Matched	Note
Barrie	1951-2019	98	-
Calgary	1956-2019	94	-
Kitchener	1985-2018	86	-
Halifax	1979-2019	99	-
Markham	1900-2018	88	-
Saskatoon	1988-2019	100	Matched In QGIS
St John's	1988-2018	100	Matched In QGIS
Vancouver	2010-2019	100	Matched In QGIS
Victoria	1985-2019	95	Matched In QGIS
Waterloo	2018	87	-
Winnipeg	1919-2019	100	Matched In QGIS
Region of Waterloo	1987-2019	94	-
Region of Durham	1974-2020	99	-

Among all classification models,XGBOOST, and among all materials, cast iron was selected to compare the results between all networks. In addition to all utilities, one global model including all networks was also created. The result of this model indicated how the accuracy of each network might affect the Global accuracy. Therefore, this could be one of the primary reasons that a uniform model can not be chosen easily.

To analyze all utilities together, the thirteen utility datasets were merged. Four new columns were added to each utility dataset before merging them in order to maintain the identity of each. The four identifying columns are as follows:

1. ***Index****: This column is a unique numerical identifier for the whole dataset.*
2. ***Utility:*** *This is the name of the utility, e.g. Vancouver.*
3. ***Abbrev:*** *This column includes the abbreviation of each utility's name, e.g. VAN for Vancouver.It is used for creating the UtilityID.*
4. ***UtilityID:*** *This is a unique pipe ID. It combines Abbrev and PipeID. The latter is available in the dataset and is specific to each utility. Its format is defined by the utility.*

After completing the data cleaning process, the most important attributes were kept for modeling. These attributes included the following:

- Material (Cast Iron, Ductile Iron, Polyvinyl Chloride (PVC), Asbestos Cement, Copper, Polyethylene, Galvanized Steel, Concrete, High-Density Poly Ethylene (HDPE), steel, and Lead)
- Diameter
- Pipe Depth
- Break Depth
- Joint Type (Ring, Lead, Welded, Universal, Collar, Gasket, Mechanical)
- Installation Date and Installation Year
- Failure Date
- Failure Month
- Replacement Year
- Soil Type (Clay, Sand, Gravel, Silt, Granular, Marsh, Muck, and Rock)
- Bedding Type (Granular, Concrete)
- Surface Type (e.g., asphalt, gravel, water, concrete)
- Protection Status
- Lining Status

- Lining Material (e.g., Cement Mortar, Epoxy)
- Protection Material (e.g., Concrete, PE)
- Lining Year
- Coating (e.g., fiberglass reinforced plastics, Concrete)
- Anode Type (Zinc, Magnesium) and Anode Status
- Casing Material (e.g., Styrofoam, Polystyrene)
- Service Type (e.g., Distribution, Transmission)
- Failure Type (e.g., circumferential, hole, crack)
- Failure Cause (e.g., corrosion, temperature change); and
- Roughness

Followings are the explanations about the most important attributes, for which a significant amount of time spent:

- ***Material***: One of the most critical factors contributing to different failures is material. Various types of pipes have different resistance against failure. For example, Cast Iron is more prone to failure (Barton et al. 2019). The manufacturing methods for different pipes have altered remarkably over the decades (Røstum, 2000). Therefore, according to previous studies, the impact of different materials should be analyzed deeply since this factor may change the pattern of failure. Almost all utilities` datasets include this attribute, which should be considered while analyzing the deterioration process. In order to clean this column, explicit categories have been defined for all utilities.

  Most utilities had this attribute in their datasets in a different way. For example, one of the datasets (Hamilton) used a different way to indicate the Cast Iron pipe (CISPIT1, CISP1, etc.). However, all of these inconsistent defined names for all cities have been changed to groups of pipes with consistent names. Moreover, in order to better understand the material itself, a full name has been used instead of an abbreviation. Finally, some datasets included different materials for one specific data point. In this case, only the first-mentioned material was taken into analysis, and the remaining were removed. In the case that material was missing from a data point, GIS was used to replace its value.

- ***Diameter***: According to previous studies, pipe break rate was found to be higher in smaller pipes than larger pipes. That is, there is an inverse relationship between Diameter and Break rate (Kettler and Goulter (1985); Andreou et al. (1987); Andreou S. A. (1986)). Being installed with thicker wall and more reliable connections, larger pipes are more resistant to ground movement and corrosion (Wengström 1993).

Nonetheless, smaller pipes are more prevalent in water distribution networks. Folkman (2018) represented that approximately 67% of pipes across Canada and the US have a diameter of less than 200 mm. Hence, it is important to know the frequency of different sizes within the network. In this study, almost all cities provided the size of water pipes. However, some utilities included the size of pipe in British Standard (inch) and some in Metric (mm). Therefore, all of these metrics have been changed to (mm) to create a consistent attribute.

- ***FailureType***: The cleaning process of this column has been the most challenging part of preprocessing. All utilities provided different formats for this attribute. Not only failure related to the nature of pipe included, but also components failures (Joint, etc.) included in this attribute. In some cases, the cause of failure is included in this part incorrectly. Therefore, all of these inconsistencies have had to be removed in order to make this attribute efficient. A few utilities provided a small group of categories for this part. However, some utilities such as St.John`s provided a complicated explanation for this attribute, including sentences, phrases, and so forth. This is the reason that made this step the most time-consuming part. In some cities, corrosion was included in the type of failure. Thus, it had to be transferred to the cause of failure. Overall, wide variety of categories were created for this attribute, which may be utilized in further analysis. Given is the small list of defined categories for Nature of Failure after the cleaning process.

    - *Circumferential*
    - *Longitudinal*
    - *Split*
    - *Crack*
    - *Blowout*
    - *Break*
    - *Leak*
    - *And …*

    There have also been some utilities that provided more than one nature of failure for different data points. These inconsistencies have also been tackled by creating mixed values, which is common among all utilities. Some of these unique categories are as follows:

    - *Circumferential/Longitudinal*
    - *Hole/Split*
    - *And …*

    As previously mentioned, there have been some failures that are not directly related to the pipe itself. These values can be found within the original datasets as follows:

Clamp, Plug, Saddle, Gasket, Elbows, Copper, Joint, Flange, Valve, Fitting, Ring, and so forth

Sleeve and Clamp have been allocated to the "Repairs" category among these elemental failures. Gasket, Joint, Flange, Fitting, Ring, Elbows, and Copper have been related to the "Joint and Fitting" category, and something like Saddle has to be defined to be in the "Not on Main" Category. Overall, there are four main categories after the cleaning process for the nature of failures.

- *Pipe-related failures (Circumferential, Crack, Longitudinal, Hole, Blowout, etc.)*
- *Joint and Fitting*
- *Repairs*
- *Not on Main*

- ***FailureCause****:* Similar to FailureType, cleansing this attribute has been a challenge, and the cleaning process has been done similar to that of FailureType. One of the main challenges in this part turns back to Corrosion, which was included in the nature of the failure. Corrosion is believed to be the cause of failure. Accordingly, corrosion was replaced and removed to failure cause for those utilities that included corrosion in nature of failure. There are some main categories for this part as follow:

  - *Accident*
  - *Corrosion*
  - *Temperature*
  - *Deterioration*
  - *Improper Installation*
  - *Improper Bedding*
  - *Differential Settlement*
  - *And …*

- ***CoatingMaterial and LiningMaterial****:* Coating and Lining are approaches that are usually employed to protect water mains against corrosion (InfraGuide, 2003; Mordak, J. and Wheeler, J., 1988). Generally, steel pipes, iron pipes, and materials more prone to failure are coated and lined (Barton et al., 2019). However, coating and lining are not generally used for PVC and PE pipes since their characteristics make them more resilient to corrosion (Barton et al., 2019). AC pipes are recommended to be coated, where surrounded by soils with a pH value of below 6.0, although concrete is generally resistant to corrosion naturally. Various types of lining and coating can be applied to different materials, such as bitumen, cement mortar, synthetic resin, PE sleeve, Epoxy, cathodic protection, and coal tar. According to what had been provided by utilities, a group of materials was defined for this attribute. These materials are consistent among all utilities.

- *CM (Cement)*
- *Epoxy*
- *Coal Tar*
- *Polyurea*
- *Concrete*
- *CIP (Cured in place)*
- *Foam*
- *Y-Jacket*
- *Urecon*
- *HDPE*
- *And …*

Most of the datasets provided by the utilities contained significant percentages of missing values that must be handled before further analyses. In the majority of cases, missing values were filled by assuming the values from adjacent pipes in GIS were applicable. Nevertheless, even after this approach, some attributes still included missing values. Replacing some of these values was straightforward. For instance, missing lining material classes were found to occur mostly for pipes with no lining, according to their lining status. Thus, the material in these cases was replaced with "Unlined." This logic was also applied to protection type, coating type, and other similar attributes.

Those pipes were removed from the dataset if general assumptions could not be made based on available GIS data and other attributes. If a particular attribute had more than 10% of missing values, they were removed entirely from the analysis. Coding in python was used to quickly detect and replace any missing values within the datasets instead of Microsoft Excel.

After data is cleaned and preprocessed, it must be prepared for the specific models that will be applied. A specific dataset was created for each type of algorithm and target, i.e., regression (rate of failure and age at first failure) and classification (probability of failure).

The rate of failure (number of breaks per year per length) is an important deterioration indicator. In order to predict the current rate of failure, i.e., of the latest break year, both the current rate of failure and the previous rate of failure were calculated. The current rate of failure was calculated based on the number of failures in the latest break year divided by the length of the pipe. The previous rate of failure, on the other hand, included all previous failures and age. This age is the same age at the current break and not necessarily the most recent age. The age at first failure was simply calculated as the difference between the first break year and the pipe install year for both targets, and only broken pipes were analyzed.

For classification models, the target variable was defined as to break status. This is a binary attribute, 0 – not broken and 1 – broken. Thus, both broken and non-broken pipes were analyzed, and the datasets were merged. Ages of broken pipes were calculated based on the

break year, while ages of non-broken pipes were calculated based on the most recent year of break data available.

3.3 Model Selection and Tools

In order to conduct this study, several tools have been applied and are described below.

- ***Python:*** In order to write the programming codes, python language was selected as it is the state-of-the-art tool for machine learning problems. The final edition of version 3 of python was installed and employed for this study.
- ***Jupyter notebook:*** The Jupyter Notebook is a well-known open-source web application that allows data scientists to create codes and make visualization. The notebook can be used for data cleaning, statistical modeling, visualization, and machine learning purposes.
- ***Scikit-Learn:*** The Scikit-learn is a free machine learning library for python programming, including different classification, regression, and clustering algorithms. This library also consists of different evaluation metrics that could help better evaluate the accuracy of the models.
- ***Matplotlib:*** Matplotlib is a graphical platform for Python and its numerical libraries, such as Numpy. This platform, along with seaborn were used for making visualizations.
- ***Seaborn:*** Seaborn is a visualization library created based upon Matplotlib and is integrated with Pandas library in python.
- ***Numpy:*** This library is integrated with Pandas and Python language, and it contains a valuable range of mathematical functions, matrices, and arrays.
- ***Pandas:*** A library built on top of python programming language that can be employed for data manipulation and perform analytical tasks.
- ***Power BI:*** This tool is a business analytic produced by Microsoft Corporation, and it can be used to create dashboards, merge and clean the datasets, using Python codes for visualizations, and many other intelligent tools that facilitate the prediction process.
- ***QGIS:*** QGIS is a free, open-source geographical platform that can be used for analyzing geospatial information. Some of the missing values within this study were handled in QGIS based on the information provided as shapefiles. The plug-in used for this step is known as “Joint to the nearest sample.”

Explanations about different algorithms have been provided in chapter 2 of this study. Based on the previous studies and trial and error, XGBOOST, ANN, Random Forest, and Logistic Regression were selected for classification models, which aimed to predict the probability of

failures. Furthermore, for regression analysis, XGBOOST, ANN, Random Forest, and ElasticNet regression were chosen.

As previously mentioned, there was not enough information for the minority class (broken pipes) in some cases (e.g., Region of Waterloo). In order to tackle this challenge, the Synthetic Minority Oversampling Technique (SMOTE) was utilized to increase the number of broken pipes artificially.

In order to evaluate classification models, the n-fold cross-validation approach was selected due to its popularity and robustness. Based on the cross-validation results, various metrics were extracted. Firstly, a confusion matrix was produced, and different metrics were created based on this matrix. Accuracy, precision, recall, and most importantly, F1-Score metrics were used for final evaluation. It is worth mentioning that for classification models, especially for imbalanced datasets, F1-Score is the best metric that can be used since it effectively considers both majority and minority classes.

For the evaluation of regression models, Root Mean Squared Error (RMSE) and R-Squared have been employed, and the performance of all models was compared to find the best predictive algorithm.

Finally, in each step and for all models, the RandomizedSearhCv from Scikit-learn was utilized for finding the best parameters. All explanations about the hyperparameters for all models and other critical information are provided in Appendix B. This includes the hyperparameters related to Random Forest (RF), Logistic Regression (LR), XGBOOST, Elastic Net Regression (ER), Artificial Neural Networks (ANN), and SMOTE.

3.4 DEPLOYMENT

Last but not least is the deployment. In order to address this step, the results were collected and compared. Then, based on this comparison, the best model was selected. For instance, for classification models, the results of XGBOOST were integrated into the dataset. Subsequently, the final dataset was merged to the GIS file and visualized the results on the map. It should be mentioned that GIS was only used for a few cities such as Saskatoon, Winnipeg, and St. John's. Using these maps would help specialists to perform well within decision-making processes.

4. AVAILABLE DATASETS

Data for this study was collected from thirteen utilities across Canada, including Vancouver, Victoria, St. John's, Saskatoon, Calgary, Kitchener, Barrie, Winnipeg, Markham, Waterloo, Region of Waterloo, Region of Durham, and Halifax. Each utility provided two datasets - either as spreadsheets or GIS shapefiles - pipe inventory and historical water main break records. The first includes all existing pipes within the network and their characteristics, whereas the second lists breaks, and the failure record includes only broken pipes. Details of the datasets are described in the following sections. Values presented were calculated based on the clean data.

All available attributes in the inventory files for each utility are shown in TABLE 4.1. All utilities provided diameter, length, material, and installation year. Thus, this attribute is common among all networks. Additionally, some attributes such as ownership, lining status, and lining material were collected by most networks. However, most of the networks across Canada did not collect some characteristics such as average soil resistivity, soil type, land type, and surface type. Calgary, for instance, is the only network that collected average soil resistivity. In addition, Victoria and Barrie are the only networks that provided hydraulic grade line (HGL) and Casing Material, respectively. In the given table, all networks can be compared based on the provided features.

Moreover, different materials were installed within various networks across Canada. The given table shows the contribution of each material to the inventory files (TABLE 4.2). As shown in the table, PVC pipes account for more than 40% of the total length of all networks, followed by cast iron and ductile iron with more than 21% contribution for each. Nonetheless, PVC is not the most popular material for all networks. For instance, based on the available datasets, cast iron accounts for the most significant portion of the network in Vancouver and Victoria, more than 45%. Alternatively, ductile iron is the most frequently installed material in Halifax, Kitchener, and St. John's. Asbestos cement is also another material that has been used in some of the networks more frequently. Therefore, it is clear that the predominant materials vary among all utilities.

Finally, another table has been provided, which shows the percentage of each material within all networks, based on the number of failures they experienced. For example, cast-iron makes up 51.67% of entire failures recorded. This shows the severity of the deterioration process of this material. Ductile iron and PVC pipes are other materials that experienced more failures than other pipes (

TABLE 4.3). For example, in Calgary, the number of failures for PVC is higher than that of cast iron and ductile iron. Therefore, based on the provided historical failures, different patterns were detected for each network. More details about each utility are provided separately in the Appendix A.

TABLE 4.1 – AVAILABLE ATTRIBUTES FOR ALL UTILITIES (INVENTORY)

Utility	Diameter	Material	Joint Type	Installation Year	Replaced Year	Replaced Status	Ownership	Length	Lining Year	Lining Status	Lining Material	Status	Coating Material	Roughness	Protection Status	Protection Year	Pipe Depth	Service Type	Bedding Type	Surface Type	Soil Type	Break Rate	Break Number	HGL	**Restrained	Casing Material	***Dead End	Average Soil Resistivity
Barrie	✓	✓	-	✓	-	-	-	✓	-	-	-	✓	-	-	✓	-	-	✓	-	-	-	-	-	-	✓	✓	-	-
Calgary	✓	✓	-	✓	-	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	✓	-	-	-	✓	✓
Halifax	✓	✓	-	✓	-	-	-	✓	-	✓	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Kitchener	✓	✓	-	✓	-	-	-	✓	✓	✓	✓	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Markham	✓	✓	-	✓	-	-	✓	✓	✓	✓	-	-	-	✓	✓	✓	✓	-	-	-	-	-	-	-	-	-	-	-
Region of Durham	✓	✓	-	✓	-	-	✓	✓	✓	✓	✓	✓	-	-	✓	✓	-	-	-	✓	-	-	-	-	-	-	-	-
Region of Waterloo	✓	✓	-	✓	-	-	✓	✓	✓	✓	✓	-	-	✓	-	-	✓	-	✓	✓	✓	-	-	-	-	-	-	-
Saskatoon	✓	✓	✓	✓	✓	✓	✓	✓	✓	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
St. John's	✓	✓	-	✓	-	-	-	✓	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Vancouver	✓	✓	-	✓	-	-	✓	✓	-	-	✓	✓	✓	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-
Victoria	✓	✓	-	✓	-	-	✓	✓	-	✓	✓	✓	-	✓	-	-	-	-	-	-	-	-	-	✓	-	-	-	-
Waterloo	✓	✓	-	✓	-	-	✓	✓	✓	✓	✓	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-
Winnipeg	✓	✓	✓	✓	-	-	✓	✓	-	-	-	✓	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Restrained means pipe is prevented from axial displacement * a pipe that is no longer used or an isolated pipe from normal water flow

TABLE 4.2 – PERCENTAGE LENGTH OF EACH MATERIAL WITHIN EACH UTILITY (INVENTORY)

Utility	Barrie	Calgary	Region of Durham	Halifax	Kitchener	Markham	Region of Waterloo	Saskatoon	St.John's	Vancouver	Victoria	Waterloo	Winnipeg	**Grand Total**
Asbestos Cement	0.13%	1.27%	2.64%	0.86%	1.07%	0.50%	7.27%	29.61%	0.15%	-	0.20%	0.34%	24.20%	6.804%
Brass	-	-	-	0.005%	-	-	-	-	-	-	-	-	-	0.000%
Cast Iron	11.50%	15.06%	12.96%	27.63%	23.13%	4.66%	7.10%	17.88%	41.89%	48.54%	45.94%	29.68%	26.08%	21.197%
Concrete	3.53%	5.06%	9.82%	5.36%	4.66%	5.82%	16.25%	-	0.65%	0.79%	-	0.03%	0.03%	4.148%
Copper	2.32%	0.33%	0.07%	0.39%	0.06%	1.62%	0.04%	2.74%	0.26%	0.46%	0.47%	0.07%	1.28%	0.766%
Cross Linked Polyethylene	0.11%	-	-	0.04%	-	0.02%	-	-	0.03%	-	-	-	-	0.009%
Ductile Iron	26.85%	20.12%	16.43%	59.36%	37.21%	18.58%	25.57%	0.10%	46.47%	44.12%	40.22%	15.12%	1.42%	21.664%
Galvanized Iron	-	-	-	-	-	-	-	-	-	-	0.08%	-	-	0.001%
Galvanized Steel	0.40%	-	-	0.001%	-	-	-	-	-	-	-	-	-	0.016%
HDPE	0.87%	0.00%	0.00%	0.17%	0.36%	0.45%	0.31%	0.07%	0.00%	0.08%	2.70%	0.43%	0.01%	0.153%
PCCP	-	-	-	-	-	-	-	-	-	-	-	-	0.00%	0.000%
Polybutylene	-	-	-	-	-	0.11%	-	-	-	-	-	-	0.01%	0.009%
Polyethylene	-	0.76%	0.09%	-	-	0.02%	0.26%	0.46%	0.02%	0.01%	0.08%	0.05%	0.10%	0.277%
PVC	54.30%	54.36%	57.99%	5.62%	31.43%	66.36%	42.92%	44.79%	10.53%	0.04%	7.83%	54.29%	46.83%	43.128%
PVCB	-	-	-	-	0.07%	-	-	-	-	-	-	-	-	0.000%
PVCF	-	0.02%	-	-	0.03%	-	-	0.05%	-	-	-	-	-	0.008%
PVCO	-	-	-	-	1.96%	-	-	-	-	-	1.44%	-	-	0.026%
Stainless Steel	-	-	-	0.06%	-	-	-	-	-	-	-	-	-	0.005%
Steel	0.00%	3.04%	-	0.52%	0.02%	1.85%	0.28%	4.30%	-	5.95%	1.04%	-	0.04%	1.789%
Grand Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**

TABLE 4.3 - PERCENTAGE OF EACH MATERIAL BASED ON THE HISTORICAL FAILURES (BREAK FILES)

Utility	Barrie	Calgary	Region of Durham	Halifax	Kitchener	Markham	Region of Waterloo	Saskatoon	StJohns	Vancouver	Victoria	Waterloo	Winnipeg	**Grand Total**
Asbestos Cement	-	1.11%	2.30%	0.38%	0.40%	0.46%	3.13%	19.40%	0.31%	0.76%	0.31%	-	15.42%	7.37%
Brass	-	-	-	0.03%	-	-	-	-	-	-	-	-	-	0.00%
Cast Iron	56.81%	33.79%	52.85%	91.60%	75.56%	44.42%	41.32%	36.54%	86.26%	90.18%	56.80%	82.56%	69.37%	51.67%
Concrete	0.55%	0.14%	1.46%	0.20%	0.25%	0.18%	3.82%	-	0.06%	0.11%	-	0.11%	0.07%	0.21%
Copper	2.38%	0.04%	0.03%	0.08%	0.05%	0.28%	-	0.14%	0.44%	0.86%	0.21%	-	0.03%	0.11%
Cross Linked Polyethylene	0.63%	-	-	-	-	-	-	-	-	-	-	-	-	0.01%
Ductile Iron	27.34%	25.81%	30.95%	7.06%	21.15%	49.37%	31.94%	0.10%	12.06%	3.67%	28.35%	14.11%	3.58%	15.73%
Galvanized	-	-	-	-	-	-	-	-	0.06%	-	-	-	-	0.00%
Galvanized Iron	-	-	-	-	-	-	-	-	-	-	0.10%	-	-	0.00%
Galvanized Steel	6.97%	-	0.23%	-	-	0.07%	-	-	-	-	-	-	-	0.11%
HDPE	0.16%	-	-	-	0.05%	0.07%	-	0.04%	0.12%	0.11%	1.86%	0.11%	0.00%	0.03%
Polybutylene	-	-	-	-	-	-	-	-	-	-	-	-	0.01%	0.00%
Polyethylene	-	1.03%	-	-	-	0.07%	-	0.24%	-	-	0.21%	-	0.00%	0.42%
PVC	5.15%	37.71%	12.18%	0.64%	2.54%	5.09%	17.36%	40.45%	0.56%	0.22%	7.63%	3.11%	11.47%	23.67%
PVCF	-	0.00%	-	-	-	-	-	0.01%	-	-	-	-	-	0.00%
PVCO	-	-	-	-	-	-	-	-	-	-	2.47%	-	-	0.02%
Steel	-	0.37%	-	-	-	-	2.43%	3.07%	0.12%	4.10%	2.06%	-	0.06%	0.65%
Grand Total	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**

TABLE 4.4 –RANGE OF NUMERICAL AND CATEGORICAL VALUES AMONG UTILITIES

	Material					Length (m)	Diameter (mm)	Installation Year	Installed after 1990	Failure Year	Failed Since 2010	Lining Material	Percentage Lined
	CI	PVC	DI	AC	Other								
Barrie	12%	54%	27%	0%	7%	0.1 - 3008	19 - 1250	1891 - 2019	73%	1951 - 2020	31%	-	-
Calgary	15%	54%	20%	1%	9%	0.05 – 993.85	12 - 3000	1900 - 2019	51%	1955 - 2019	15%	-	-
Halifax	28%	6%	59%	1%	7%	0.03 - 3620	19 - 1500	1856 - 2019	41%	1956 - 2019	32%	CM, Polyurea, Unlined	48.25%
Kitchener	23%	32%	37%	1%	7%	0.01 – 83.46	25 - 1200	1887 - 2018	47%	1985 - 2018	42%	CM, EPOXY, Unlined	0.34%
Markham	5%	66%	19%	1%	10%	0.82 – 3779	25 - 1800	1938 - 2019	63%	1979 - 2019	11%	-	12.83%
Region of Durham	13%	58%	16%	3%	10%	0.15 – 4169.73	25 - 2100	1900 - 2020	52%	1972 - 2020	24%	CIP, CM, Unlined	14.30%
Region of Waterloo	7%	43%	26%	7%	17%	0.06 – 6977	38 - 1200	1850 - 2019	48%	1987 - 2019	55%	CM, Unlined	32.82%
Saskatoon	18%	45%	0%	30%	8%	0.2 - 1581	25 - 1350	1906 - 2019	39%	1958 - 2019	17%	CIP, HDPE, and PE	1.81%
St. John's	42%	11%	46%	0%	1%	-	12 - 1400	1892 - 2017	36%	1988 - 2018	27%	-	-
Vancouver	49%	0%	44%	0%	7%	0.09 - 2210	20 - 1950	1892 - 2020	37%	2009 - 2020	99%	-	-
Victoria	46%	8%	40%	0%	6%	0.15 - 716	19 - 990	1888 - 2016	24%	1985 - 2019	31%	CM, EPOXY, HDPE, Unlined	4.70%
Waterloo	30%	54%	15%	0%	1%	0.09 – 644	25 - 450	1850 - 2018	41%	2000 - 2020	55%	CM, Epoxy, HDPE, Unlined	11.41%
Winnipeg	26%	47%	1%	24%	1%	0.01 – 996.59	19 - 1050	1882 - 2020	47%	1919 - 2019	20%	-	-

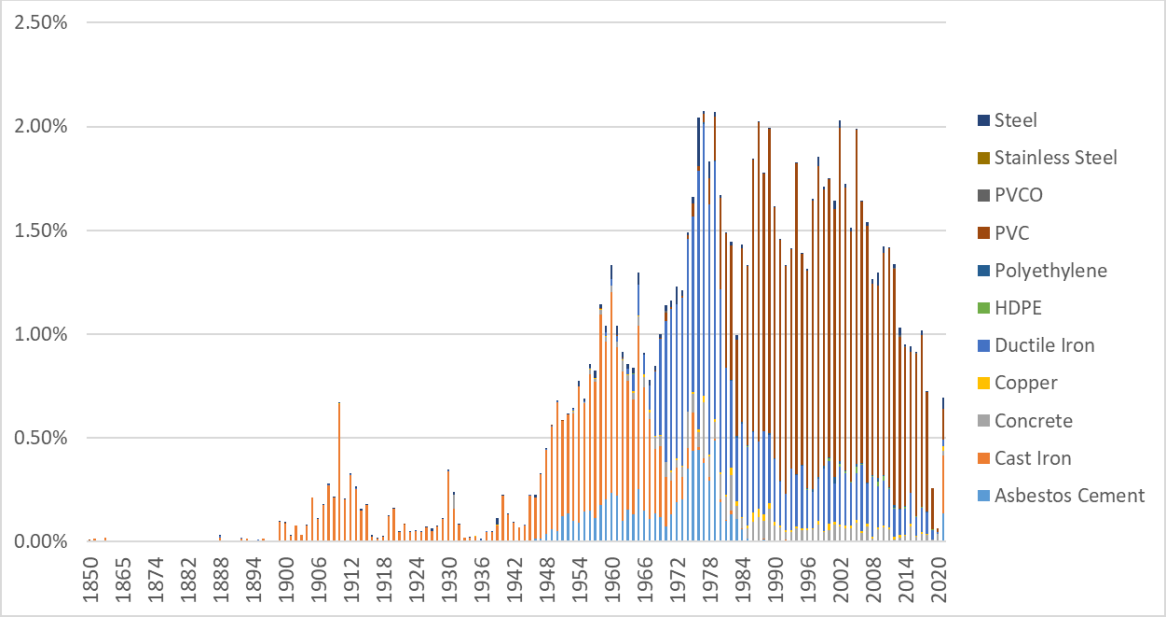


FIGURE 4.1 – PERCENTAGE OF PIPE INSTALLED IN DIFFERENT YEARS BASED ON THE AVAILABLE INFORMATION

The given figure indicates the percentage of each material installed in various years within all utilities (Figure 4.1). As seen in the graph, cast iron was the predominant type of material from the beginning of the pipe installation until the early 70s. From then, however, ductile iron and PVC pipes have become more prevalent, with PVC pipes being installed more frequently in recent years. Based on the available information, PVC pipes are more prone to failure during the early stage of their lives. Considering this pattern and performing an in-depth analysis is required in order to better understand the failure pattern related to this pipe. In doing so, practitioners and asset management planners would be able to prepare short or long-term replacement and maintenance practices regarding PVC pipes.

The given chart shows the frequency of various sizes within provided information (Figure 4.2). Apparently, smaller pipes contribute to a significant proportion of existing networks. 150-mm pipes account for almost one-third of all pipes. This size is then followed by 200-mm and 300-mm mains, with 24% and 16% contribution, respectively.

Furthermore, another bar chart has been created, which shows an increase in the rate of failures among the entire available datasets within this project (Figure 4.3). As mentioned in previous studies, the rate of failure for a certain type of pipe has undergone a significant increase. This pattern emphasizes the importance of prediction in order to ameliorate the current condition. This could be achieved by combining predictive analysis and managerial practices, which facilitate the maintenance of the precious water networks.

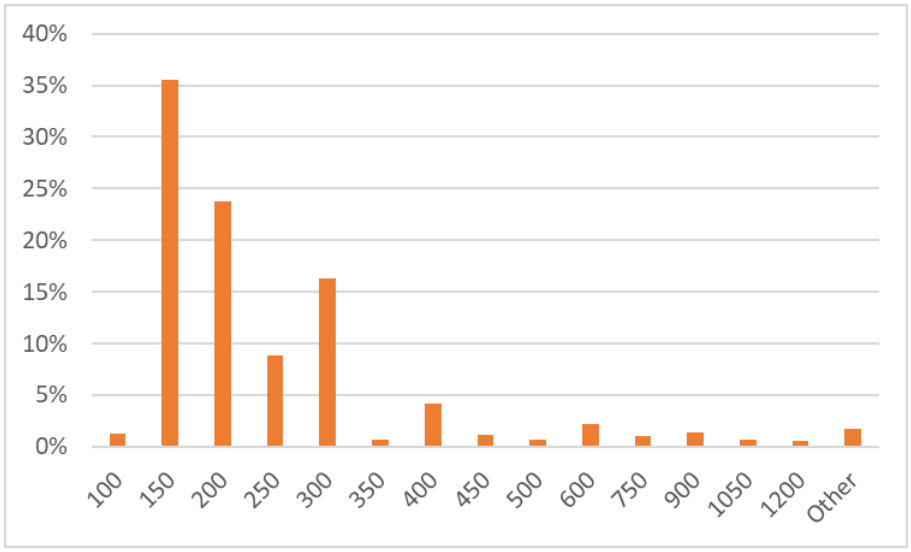


FIGURE 4.2 – PERCENTAGE OF EACH SIZE WITHIN THE ENTIRE DATASETS FOR ALL UTILITIES

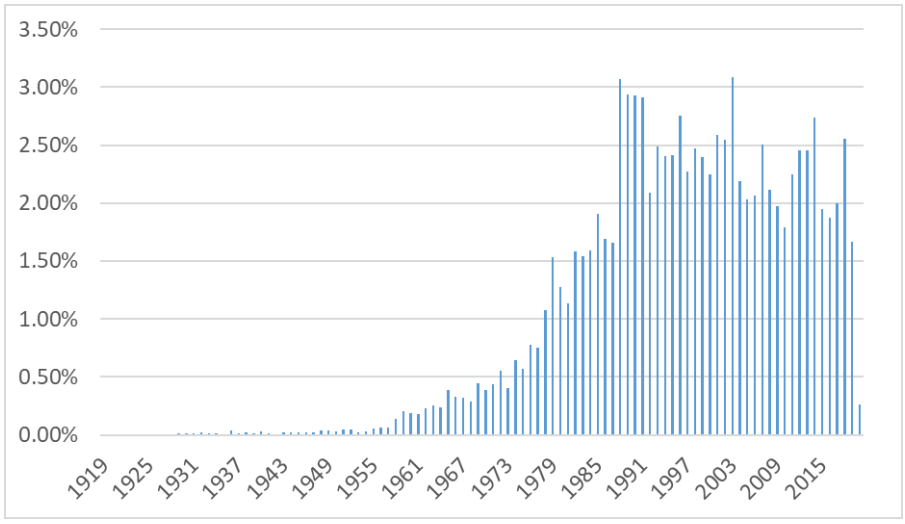


FIGURE 4.3 – PERCENTAGE OF FAILURES IN DIFFERENT YEARS BASED ON THE AVAILABLE INFORMATION

As discussed before, one of the primary challenges in this study was the imbalanced nature of some of the datasets. TABLE 4.5 shows the number of available pipes and break records for each city, as well as the ratio between broken and non-broken pipes. The table indicates that Waterloo, Region of Waterloo, and Vancouver are among the utilities with fewer failure records provided. For instance, Vancouver has a significantly imbalanced format with a ratio of 1.24% between non-broken and broken pipes. The effect of this ratio can be seen in the results of classification models. The utility with a lower ratio represented exceptionally low accuracy.

TABLE 4.5 – NUMBER OF PIPES AND BREAKS FOR EACH UTILITY

Utility	Number of Pipes	Number of Breaks	Number of Broken Pipes	The ratio of Broken to Non-Broken
Saskatoon	35,630	14,152	5,103	(13:100)
Winnipeg	114,824	26,631	10,056	(9:100)
Kitchener	14,561	2,348	987	(7:100)
Markham	10,802	2,926	628	(6:100)
Waterloo	7,565	925	446	(6:100)
Region of Waterloo	5,139	292	105	(2:100)
Region of Durham	22,414	6,578	2,026	(11:100)
Calgary	55,561	36,396	9,180	(9:100)
Vancouver	67,522	927	835	(9:1000)
Victoria	3,319	977	836	(15:100)
Halifax	14,436	6,381	5,936	(16:100)
St. John's	8,983	1,626	1,460	(10:100)
Barrie	6,522	1,297	1,240	(6:100)

TABLE 4.6 – RATIO OF BROKEN PIPES TO NON-BROKEN PIPES IN CLASSIFICATION DATASETS

Utility	All Materials	CI	HDPE	PVC	DI	AC	ST	CON	CO
Barrie	0.06	1.01	-	0.01	0.07	-	-	-	-
Calgary	0.09	0.51	-	0	0.15	0.24	0.02	0.02	-
Halifax	0.16	0.67	-	0.03	0.05	-	-	0.03	-
Kitchener	0.07	0.30	-	0.00	0.05	-	-	-	-
Markham	0.06	1.50	-	0.01	0.28	-	-	0.01	-
Region of Durham	0.10	0.81	-	0.01	0.25	0.19	-	0.01	-
Region of Waterloo	0.02	0.10	-	-	-	-	-	-	-
Saskatoon	0.13	0.47	-	0.01	-	0.20	0.26	-	0.03
St. John`s	0.10	0.24	-	0.00	0.03	-	-	-	-
Vancouver	0.01	0.02	-	-	0.00	-	-	-	-
Victoria	0.15	0.27	0.01	0.08	0.07	-	-	-	-
Waterloo	0.06	0.19	-	0.00	0.06	-	-	-	-
Winnipeg	0.09	0.43	-	0.00	0.37	0.00	-	-	-

The given tables compare the range of available numerical attributes based on all the available information. It should be mentioned that the first table shows descriptive statistics for inventory files, and the following table indicates this information for historical failures TABLE 4.7, TABLE 4.8.

TABLE 4.7 – DESCRIPTIVE STATISTICS FOR ALL UTILITIES INCLUDING NUMERICAL ATTRIBUTES - INVENTORY

Descriptive Statistics	Diameter (mm)	Length (m)	InstallYear	Average SoilResistivity (ohm-meter)	Roughness (μ)	PipeDepth (m)
Count of Values	360514	362882	354886	16179	12217	8471
mean	223.48	53.25	-	2489.15	102.70	1.13
std	128.95	97.46	-	1480.25	34.77	0.50
min	12	0	1850	597	10	0
25%	150	2.5	1967	1636	57	0.85
50%	200	14.20	1987	2092	120	0.85
75%	250	76.70	2002	2843	120	1.75
max	3000	6977.83	2020	25000	178	4.20

TABLE 4.8 - DESCRIPTIVE STATISTICS FOR ALL UTILITIES INCLUDING NUMERICAL ATTRIBUTES - BREAK

Descriptive Statistics	Failure Year	Pipe Depth (m)	Diameter (mm)	Lining Year	Protection Year	Length (m)
count	77759	7526	29463	1858	2141	8218
mean	-	2.52	196.31	-	-	171.05
std	-	1.76	92.93	-	-	145.93
min	1919	0.02	13	1970	1985	0.041
0.25	1988	1.83	150	1980	2009	83.49
0.5	1997	2.74	150	2003	2014	134.85
0.75	2008	2.9	200	2009	2017	226.70
max	2020	54.00	1950	2019	2019	2105.96

5. RESULTS

This section of this study presents the final results of the deterioration models and compares them across cities.

5.1 CLASSIFICATION

Four classification machine learning models were applied to predict water main probability of failure: random forest, XGBOOST, logistic regression, and artificial neural networks. The cross-validation method was used for the evaluation process. Accordingly, 80% of the dataset was selected for training and validation and 20% for testing. The models were evaluated based on 5-fold cross-validation, and the average of all folds was considered the final output. Results of all models are compared for each utility in TABLE 5.1. These models were created for all materials. The attributes included in each model differ by utility, depending on data availability.

Overall, from the table, it is clear that XGBOOST algorithm provided better performance for nine utilities. Due to its complex learning process, this model can detect the most intricate patterns within the datasets. XGBOOST is the developed and more regularized version of the gradient boosting method (Swamynathan, 2019). This method, specifically, employs gradient descent to build sequential decision trees that decline residuals (Snider and McBean, 2018). For two utilities, the random forest also performed similarly to XGBOOST algorithm in the Region of Durham and Halifax. With an F1-Score of 87%, Saskatoon achieved the highest accuracy among all utilities.

Furthermore, ANN also indicated relatively desirable results for some of the cities. For instance, the ANN model for Vancouver and the Region of Waterloo performed more accurately than other algorithms, with F1-Score of 26% and 19%, respectively. The accuracy of the models for these utilities is significantly low due to their imbalanced data. Moreover, other networks such as St. John's, Waterloo, and Victoria did not show desirable results, with an F1-Score of 64%, 55%, and 55%, respectively. For Victoria and St. John's, the lower accuracy was not related to the ratio of datasets. This was because the algorithms could not find appropriate patterns for predictions, requiring more analytical processes to improve the models. Finally, logistic regression, although in some cases powerful to detect broken pipes, did not provide a desirable accuracy.

These algorithms have been created and learned based on different hyperparameters, optimized for each utility using RandomizedSearchCV and GridSearchCV. A detailed explanation of this process and the resulting hyperparameters is described in Appendix B and K. Nonetheless, some algorithms, such as ANN, may provide better accuracy with further tuning.

TABLE 5.1 –RESULTS OF CLASSIFICATION MODELS FOR ALL UTILITIES AND ALL MATERIALS

Utility	Random Forest			XGBOOST			LR			ANN		
	A	F1	R	A	F1	R	A	F1	R	A	F1	R
Saskatoon	*96%*	*80%*	*70%*	***97%***	***87%***	***81%***	*95%*	*76%*	*67%*	*97%*	*85%*	*79%*
Winnipeg	*96%*	*70%*	*58%*	***86%***	***75%***	***66%***	*94%*	*54%*	*43%*	*96%*	*73%*	*65%*
Kitchener	*96%*	*64%*	*48%*	*96%*	*63%*	*51%*	*96%*	*59%*	*44%*	***96%***	***66%***	***52%***
Waterloo	*95%*	*46%*	*31%*	***95%***	***55%***	***42%***	*94%*	*36%*	*24%*	*95%*	*51%*	*38%*
***Region of Waterloo**	*98%*	*8%*	*4%*	*97%*	*7%*	*4%*	*71%*	*11%*	*70%*	***97%***	***19%***	***13%***
Region of Durham	***98%***	***85%***	***75%***	***97%***	***85%***	***78%***	*97%*	*81%*	*71%*	*87%*	*83%*	*77%*
Calgary	*98%*	*86%*	*79%*	***98%***	***88%***	***82%***	*97%*	*81%*	*71%*	*98%*	*87%*	*81%*
***Vancouver**	*99%*	*16%*	*9%*	***99%***	***26%***	***16%***	*67%*	*5%*	***91%***	***99%***	***26%***	***16%***
Victoria	*90%*	*52%*	*35%*	***90%***	***55%***	***40%***	*88%*	*45%*	*32%*	*88%*	*35%*	*21%*
St. John's	*93%*	*52%*	*37%*	***94%***	***64%***	***52%***	*78%*	*38%*	***72%***	*93%*	*52%*	*41%*
Halifax	*95%*	*79%*	*69%*	***95%***	***79%***	***72%***	*94%*	*75%*	*68%*	***95%***	***79%***	***72%***
Barrie	*97%*	*72%*	*62%*	*97%*	*71%*	*66%*	*97%*	*68%*	*57%*	***97%***	***73%***	***66%***
Markham	*98%*	*85%*	*76%*	*98%*	*85%*	*77%*	*98%*	*81%*	*72%*	***98%***	***86%***	***80%***

* Imbalanced Datasets (Note: Results in bold are the best performing models for each utility) ***A: Accuracy, R: Recall, F1: F1-Score***

All models were developed with three approaches: (1) predicting deterioration of all materials, (2) predicting deterioration of all materials and applying SMOTE, and (3) predicting deterioration of only cast-iron pipes. TABLE 5.2 compares the results for these three approaches applied to XGBOOST for each utility. Comparing all provided results indicates how each model could perform differently among entire networks and various materials. Moreover, the provided results manifest how an imbalanced dataset could considerably affect the overall performance of a classification model. As can be seen with using SMOTE, no significant enhancement was found in this study, except for the Region of Waterloo. The F1-Score for this network increased from 7% to 28%. Appealing is the effect of creating categories by materials. Almost in all utilities, the performance of XGBOOST increased from AM category to the cast iron category. This indicates that the material can be used for creating homogenous groups of pipes.

The results for other materials are provided in Appendix H. If sufficient information is available for a specific type of pipe, it is easier for the models to find a pattern that leads to higher performance. The availability of information should be sufficient for both broken and non-broken pipes. For instance, in the Saskatoon network, since the historical failure for all pipes was relatively adequate, XGBOOST achieved an F1-Score of 81%. On the other hand, in some utilities where data for a predominant type of material is not enough, the result was unsatisfactory. This emphasizes the importance of data availability and data collection for future predictions. Logistic regression was the model that showed its power to predict broken pipes where data is not available. Nevertheless, this model was not able to detect non-broken pipes efficiently. Therefore, the overall performance of this model did not indicate enough reliability.

TABLE 5.2 –RESULTS OF XGBOOST MODELS FOR ALL UTILITIES UNDER THREE APPROACHES

Utility	AM			SMOTE			Cast Iron		
	A	F1	R	A	F1	R	A	F1	R
Saskatoon	*97%*	*87%*	*81%*	*96%*	*82%*	*84%*	**94%**	**89%**	***86%***
Winnipeg	*86%*	*75%*	*66%*	*94%*	*68%*	***84%***	**87%**	**75%**	***69%***
Kitchener	*96%*	*63%*	*51%*	*90%*	*54%*	***80%***	**91%**	**78%**	***73%***
Waterloo	*95%*	*55%*	*42%*	*94%*	*61%*	***69%***	**91%**	**67%**	***58%***
***Region of Waterloo**	*97%*	*7%*	*4%*	*96%*	*28%*	*30%*	**87%**	**31%**	***22%***
Region of Durham	*97%*	*85%*	*78%*	*97%*	*82%*	*82%*	**91%**	**89%**	***84%***
Calgary	*98%*	*88%*	*82%*	*94%*	*81%*	***89%***	**93%**	**90%**	***88%***
***Vancouver**	*99%*	*26%*	*16%*	*91%*	*12%*	***65%***	**99%**	**28%**	***17%***
Victoria	*90%*	*55%*	*40%*	*89%*	*57%*	*52%*	87%	70%	***56%***
St. John's	*94%*	*64%*	*52%*	*91%*	*58%*	***69%***	**90%**	**72%**	***68%***
Halifax	*95%*	*79%*	*72%*	*93%*	*78%*	*84%*	**87%**	**82%**	***81%***
Barrie	*97%*	*71%*	*66%*	*95%*	*67%*	*79%*	**82%**	**83%**	***82%***
Markham	*98%*	*85%*	*77%*	*98%*	*80%*	*79%*	**94%**	**95%**	***90%***

* Imbalanced Dataset (NOTE: RESULTS IN BOLD ARE THE BEST PERFORMING MODELS FOR EACH UTILITY) ***A: Accuracy, R: Recall, F1: F1-Score***

In another step of the classification model, a Global model was created based on all available information. For this Global method, three primary attributes which were common among all utilities were chosen; length, diameter, and age. For this step, only ***cast iron (CI)*** material was chosen for the analysis.

TABLE 5.3 – COMPARISON OF GLOBAL AND UTILITY SPECIFIC XGBOOST MODELS FOR CAST IRON PIPES

System	XGBOOST (Global Application)			XGBOOST (Each network Separately with length, age, diameter)			XGBOOST (Each network Separately)		
	A	**F1**	**R**	**A**	**F1**	**R**	**A**	**F1**	**R**
****Global**	***90%***	***72%***	***61%***	*90%*	*72%*	*61%*	*90%*	*72%*	*61%*
Saskatoon	***97%***	***95%***	***92%***	*90%*	*83%*	*78%*	*94%*	*89%*	*86%*
Winnipeg	***91%***	***83%***	***78%***	*87%*	*76%*	*68%*	*87%*	*75%*	*69%*
Kitchener	***98%***	***94%***	***90%***	*90%*	*77%*	*72%*	*91%*	*78%*	*73%*
Waterloo	***98%***	***93%***	***88%***	*90%*	*65%*	*58%*	*91%*	*67%*	*58%*
***Region of Waterloo**	***100%***	***98%***	***97%***	*89%*	*43%*	*33%*	*87%*	*31%*	*22%*
Region of Durham	***99%***	***99%***	***98%***	*91%*	*88%*	*84%*	*91%*	*89%*	*84%*
Calgary	***97%***	***95%***	***92%***	*94%*	*90%*	*87%*	*93%*	*90%*	*88%*
***Vancouver**	***99%***	***48%***	***32%***	*98%*	*17%*	*10%*	*99%*	*28%*	*17%*
Victoria	***99%***	***98%***	***96%***	*85%*	*65%*	*59%*	*87%*	*70%*	*56%*
St. John's	***97%***	***93%***	***88%***	*89%*	*70%*	*66%*	*90%*	*72%*	*68%*
Halifax	***98%***	***97%***	***96%***	*86%*	*81%*	*80%*	*87%*	*82%*	*81%*
Barrie	***99%***	***99%***	***98%***	*82%*	*83%*	*82%*	*82%*	*83%*	*82%*
Markham	***100%***	***99%***	***100%***	*94%*	*95%*	*90%*	*94%*	*95%*	*90%*

* Imbalanced Dataset (NOTE: RESULTS IN BOLD ARE THE BEST PERFORMING MODELS FOR EACH UTILITY) ** Global model includes age, diameter, material, and target (broken or non-broken)

TABLE 5.3 compares the result of the Global model and its application to other utilities with the application of XGBOOST to each network separately. There are two primary columns within the table: Global Application and Each network separately. The Global model was created in the former, and the outcome was tested on the other utilities. As shown in the table, the Global

model has an accuracy of 90% and an F1-Score of 72%. This model was then applied to other networks, and results can be noticed in the table. It should be noted that the Global model was created and compared to the result of the application of XGBOOST to each network separately, which means that the models were created for each network separately.

According to the results, creating one Global model based on the specific material (Cast iron in this case) for all datasets and applying it to another network – separately - may be a practical process. For instance, a model created only for Saskatoon for cast iron pipes has an F1-Score of 89%. However, when the Global model was applied to the Saskatoon network, the F1-Score increased to 95%. This pattern can be noticed for all utilities within the table. In particular, there is a significant improvement from 31% to 98% for the Region of Waterloo.

Vancouver, another utility with an imbalanced dataset, underwent an increase in F1-Score when applying the Global model. It should be noted that although the improvement can be seen, all models should be meticulously examined in order to find out about any overfitting and underfitting within the models.

Overall, it seems that the number of available data points plays an important role in prediction. Since the available information for the Global model is sufficient, the accuracy increased due to the efficient training process.

Figure 5.1 represents the contribution of each attribute to the learning process of XGBOOST, for Cast Iron pipes. As seen in the given bar chart, age at first failure plays an important role with a contribution of almost 78% to the prediction. On the other hand, length and diameter with 12% and 10% are not as significant as age in the global model. As previously discussed, age's importance has been investigated in many studies before. Accordingly, the result of this study emphasizes that age could be considered an important factor for future prediction.

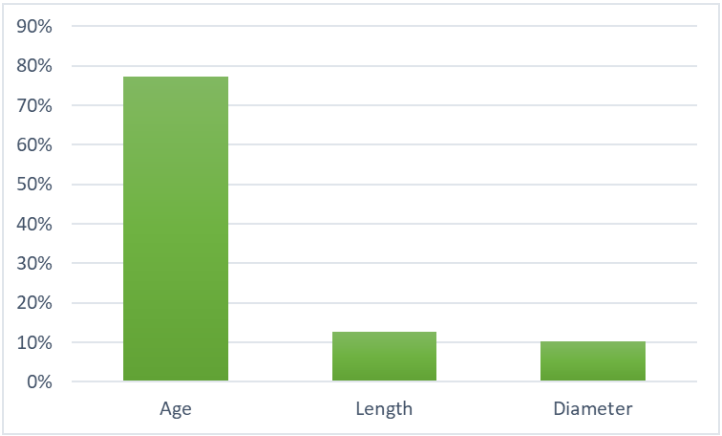


FIGURE 5.1 – CONTRIBUTION OF EACH ATTRIBUTE TO CREATING THE GLOBAL MODEL

In addition to analyzing all materials and cast iron pipes within different steps, classification models were also created for other types of materials. Overall, in the case where a failure record is available for a specific pipe, the machine learning algorithms are better able to find a pattern. However, in the case where data is not collected or is not available sufficiently, a

model is less likely to detect a specific pattern that can be used for future prediction. In this study, for instance, historical failure information related to the cast iron, ductile iron, and asbestos cement pipes is relatively adequate. Thus, the algorithms provided better results. On the other hand, some pipes such as PVC and HDPE experienced less rate of failures compared to others. This is the reason why machine learning models find the prediction for this type of materials less efficient, with having lower performance. The comparison between the results of different materials is provided in appendix H.

Furthermore, these algorithms predict the probability of failure of each pipe and can thus also be visualized in GIS. Mapping the results facilitates their understanding and communication. It can also improve the transparency of the machine learning outputs, which can sometimes be seen as a "black-box." Example maps are shown below for Saskatoon. Figure 5.2 shows the host spot locations within the Saskatoon water main network. The most critical areas can be detected and managed more efficiently based on this map. Another map, on the other hand, indicates each segment individually (Figure 5.3).

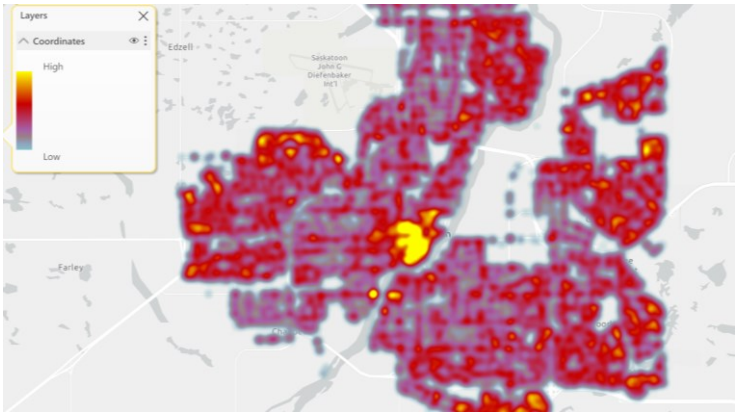


FIGURE 5.2 – MAP OF PROBABILITY OF FAILURE HOT SPOTS (SASKATOON)

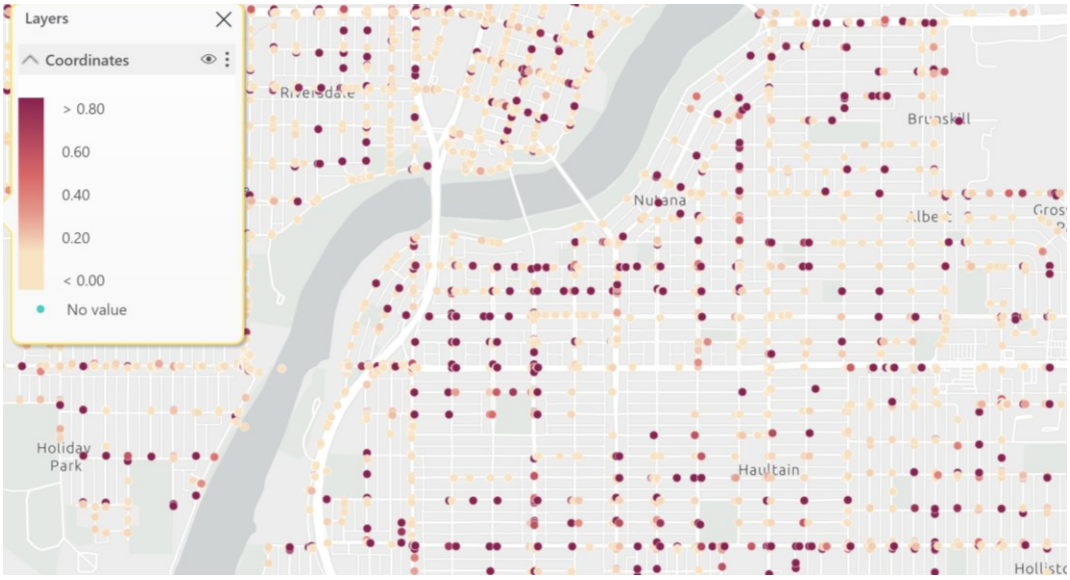


FIGURE 5.3 – MAP OF PROBABILITY OF FAILURE (SASKATOON)

Overall, the material has the most significant influence on the prediction. However, the type of material may vary between different utilities. For instance, PVC pipes may have the highest effect in one network (Saskatoon, Winnipeg, Waterloo, Region of Waterloo), but other utility cast iron pipes (Region of Durham, Calgary, Halifax, Barrie). Age, length, and lining status were also found to have a considerable impact. For instance, in Kitchener, age and length were found to be the most influential attributes. Alternatively, in Markham, lining age with more than 50% contribution was found to have the highest impact on the prediction process. The effect of each attribute on the learning and prediction processes can be found in Appendix C.

5.2 REGRESSION

This section compares the results achieved for all networks for regression analysis. Two targets were defined for regression models: age at first failure and the current rate of failures. Four algorithms were applied for each target: random forest, XGBOOST, ANN, and ElasticNet. These models were then compared using different regression metrics – RMSE, R-Squared - and related results are provided in more detail. Models were developed to predict the deterioration of all materials or specifically cast iron. In some cases, there are not sufficient cast iron pipes for creating the model. In these cases, the regression models were only applied to all materials. It should also be noted that the regression models focus strictly on broken pipes. Thus, imbalanced data is not an issue. Cross-validation was also applied here for evaluation. For this purpose, 75% of the entire dataset was allocated to the training and validation split and 25% to the testing split. Moreover, similar to that of classification, 5-fold cross-validation was chosen for evaluating the models.

5.2.1 Age at First Failure

TABLE 5.4 compares the results achieved from regression analysis for the prediction of age at the first failure. As previously discussed, the results for regression problems are not desirable enough. Overall, the random forest algorithm was the best model for the majority of the networks. For instance, for the Region of Waterloo, random forest with an RMSE of 14.46 and an R-Squared of 0.67 performed better than other models. Alternatively, in Winnipeg, results for XGBOOST and random forest were relatively close. XGBOOST also provided the best result for Waterloo utility With an R-Squared of 0.42.

TABLE 5.4 – RESULTS OF REGRESSION MODELS PREDICTING AGE AT FIRST FAILURE FOR ALL UTILITIES

Utility	ElasticNet		Random Forest		XGBOOST		ANN	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
Saskatoon	*17.85*	*0.47*	***17.82***	***0.48***	*17.93*	*0.47*	*17.86*	*0.47*
Winnipeg	*19.82*	*0.45*	***19.48***	***0.47***	***19.55***	***0.47***	*21.86*	*0.33*
Kitchener	*11.33*	*0.38*	***11.29***	***0.39***	*11.68*	*0.34*	*11.35*	*0.38*
Waterloo	*11.79*	*0.16*	*10.57*	*0.32*	***9.82***	***0.42***	*11.75*	*0.164*
Region of Waterloo	*21.47*	*0.29*	***14.46***	***0.67***	*17.42*	*0.53*	*22.43*	*0.22*
Region of Durham	*15.07*	*0.18*	***14.91***	***0.20***	*15.18*	*0.17*	*15.11*	*0.18*
Calgary	*16.53*	*0.12*	***16.53***	***0.12***	*16.66*	*0.11*	*16.57*	*0.12*
Vancouver	*16.18*	*0.32*	*15.67*	*0.36*	*16.39*	*0.304*	***15.38***	***0.39***
Victoria	*18.47*	*0.30*	***17.26***	***0.38***	*17.96*	*0.33*	*18.60*	*0.28*
St. John's	*18.13*	*0.23*	***17.60***	***0.27***	*178.9*	*0.24*	*12.92*	-
Halifax	*18.45*	*0.20*	***17.15***	***0.31***	*17.50*	*0.28*	*19.84*	*0.07*
Barrie	*14.79*	*0.29*	***14.22***	***0.35***	*16.01*	*0.17*	*14.46*	*0.32*
Markham	*11.18*	*0.15*	***11.05***	***0.17***	*11.64*	*0.08*	*12.29*	-

NOTE: RESULTS IN BOLD ARE THE BEST PERFORMING MODELS FOR EACH UTILITY

More importantly, it should be noted that comparing the results between all cities is not an appropriate method since each utility has a different material composition and average age at first failure. Furthermore, each utility is affected by different factors, which makes the comparison more challenging. In order to better understand how well each model fits the dataset, the given bar chart is prepared, which shows the average age at failure for different utilities (Figure 5.4).

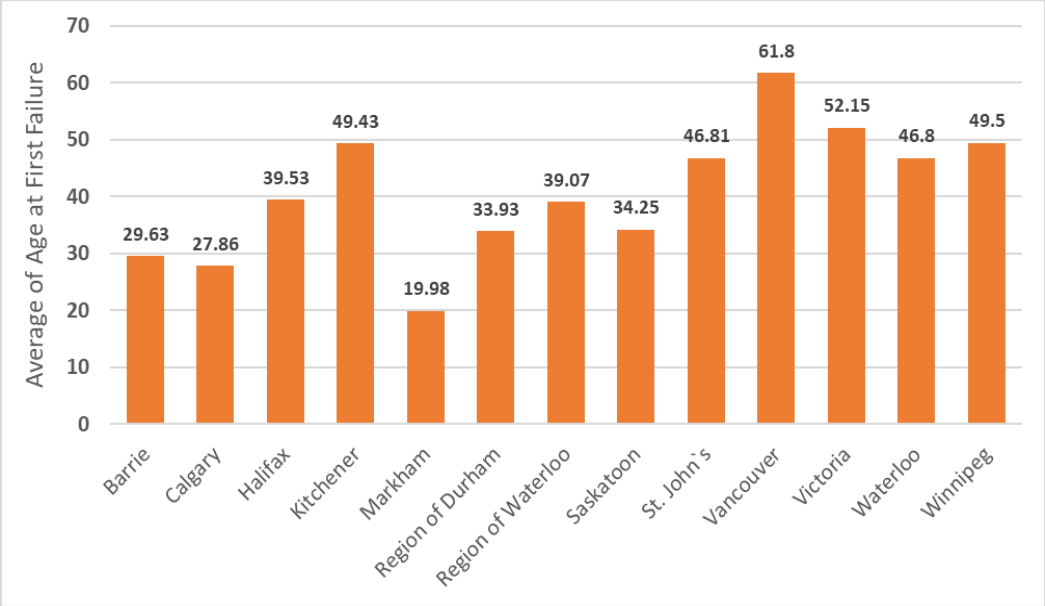


FIGURE 5.4 – AVERAGE AGE AT FIRST FAILURE FOR ALL UTILITIES (ALL MATERIALS)

In order to further evaluate the difference among materials, box plots of age at first failure were also created for cast iron, PVC, and ductile iron (Figure 5.5, Figure 5.6, Figure 5.7).

The average age at first failure for cast-iron pipes ranges from 20 for Markham to around 60 for the Region of Waterloo and Winnipeg. This value is generally lower for ductile iron pipes. It is also lower for PVC, but the variance is markedly greater. This is likely due to the fact that PVC has become more popular in recent decades, and most PVC pipes have not yet reached the end of their useful life.

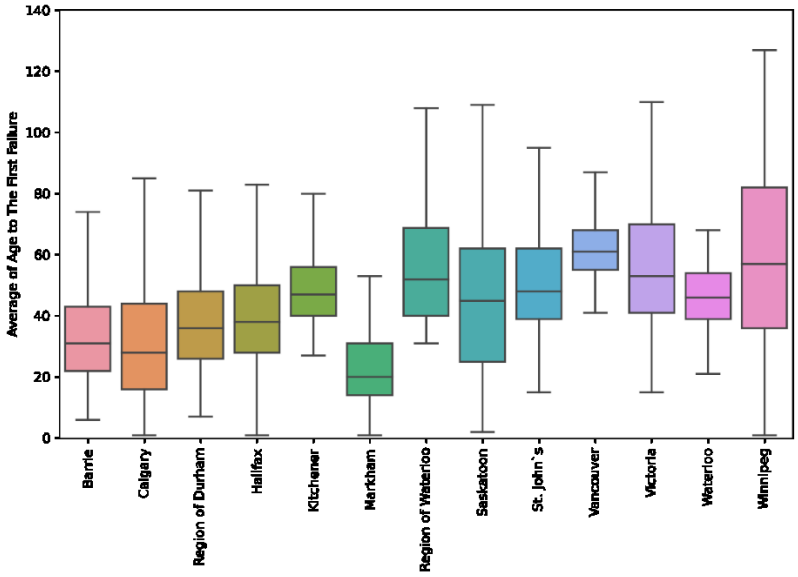


FIGURE 5.5 – AVERAGE AGE AT FIRST FAILURE IN ALL UTILITIES (CAST IRON)

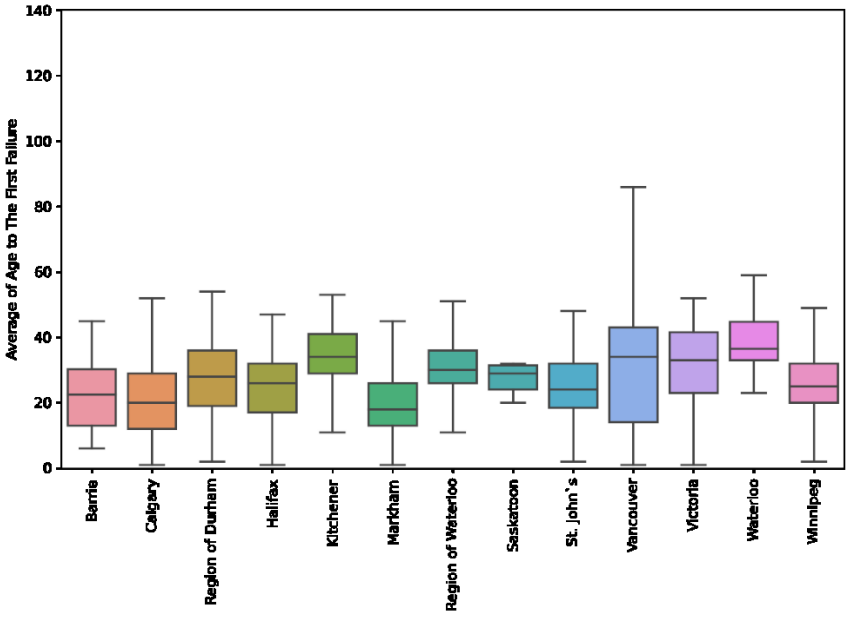


FIGURE 5.6 - AVERAGE AGE AT FIRST FAILURE IN ALL UTILITIES (DUCTILE IRON)

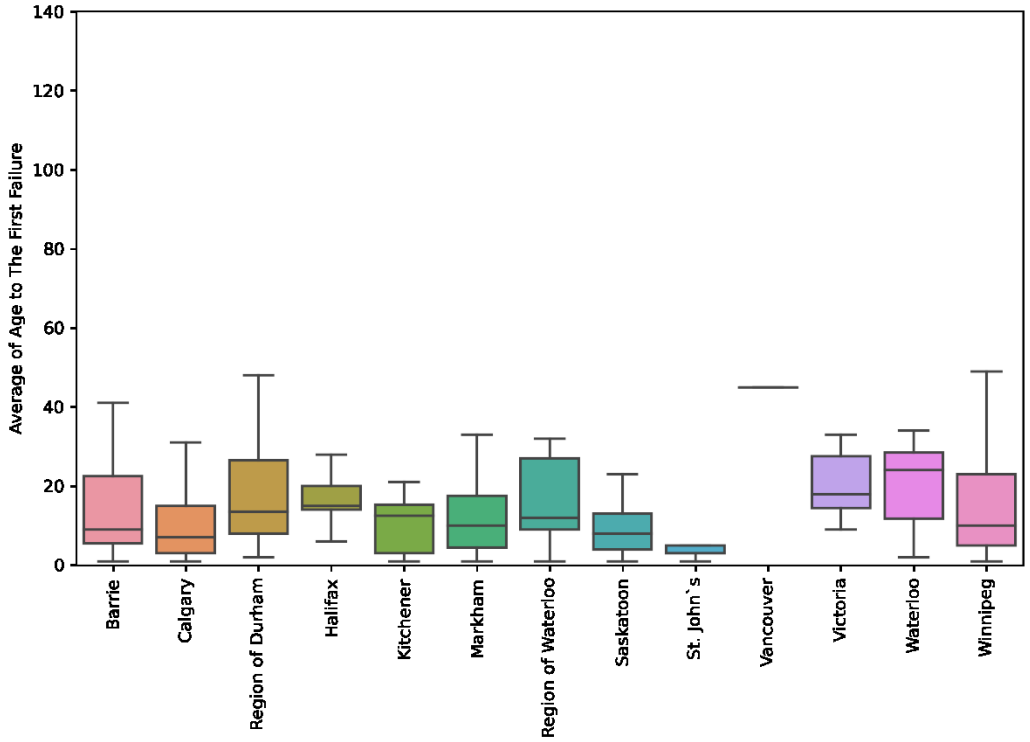


FIGURE 5.7 – AVERAGE AGE AT FIRST FAILURE IN ALL UTILITIES (PVC)

All regression models are able to provide the influence of each attribute on the prediction process. In this study, each feature's importance was provided based on the result of the random forest as the best model for the prediction of number age at first failure. Different attributes have different contributions to learning and prediction in various networks. Given bar charts in Appendix C provide more information regarding the impact of different input variables. In most networks, the material was found to be highly influential.

Nonetheless, joint type, length, and diameter are also among the most critical attributes. For instance, in Saskatoon, the lead joint has the highest contribution to age at first failure, followed by cast iron material. Or, in Markham and Barrie, length was found to have the highest contribution to learning and prediction. Lining age has a significant effect on the Waterloo network. Victoria and Halifax indicated that diameter could be another influential attribute in some of the networks. As seen, different attributes should be scrutinized for each network separately.

5.2.2 CURRENT RATE OF FAILURE

The table below compares various regression models used to predict the current rate of failure, considering different utilities (TABLE 5.5). For most utilities, XGBOOST indicated a better performance compared to other models and showed a desirable accuracy. The random forest

also performed relatively well in most cases, and for some cities was the best predictive model. However, a few critical points should be mentioned here. It is not justifiable to compare all utilities based on the given scores since each network behaves differently and should be analyzed individually. For instance, based on the available information, Winnipeg's average current rate of failure is 0.082 per meter for the entire network. The RSME score for Winnipeg is 0.045 based on the XGBOOST result. Comparing RMSE and the average current rate of failure shows that the model is relatively acceptable since RMSE is almost half of the average rate of failure.

TABLE 5.5 - RESULTS OF REGRESSION MODELS PREDICTING RATE OF FAILURE FOR ALL UTILITIES

Utility	ElasticNet		Random Forest		XGBOOST		ANN	
	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
Saskatoon	*0.125*	*0.176*	***0.028***	***0.959***	***0.028***	***0.959***	*0.054*	*0.848*
Winnipeg	*0.443*	*0.033*	*0.138*	*0.906*	***0.045***	***0.990***	*0.703*	-
Kitchener	*1.067*	*0.293*	*0.519*	*0.833*	***0.383***	***0.909***	*0.396*	*0.902*
Waterloo	*0.059*	*0.119*	*0.040*	*0.580*	***0.030***	***0.782***	*0.047*	*0.443*
Region of Waterloo	*0.260*	*0.074*	***0.114***	***0.821***	*0.120*	*0.801*	*0.246*	*0.169*
Region of Durham	*0.024*	*0.092*	*0.012*	*0.783*	***0.012***	***0.784***	*0.023*	*0.145*
Calgary	*0.020*	*0.152*	*0.007*	*0.903*	***0.005***	***0.957***	*0.018*	*0.326*
Vancouver	*0.245*	*0.126*	***0.034***	***0.983***	*0.036*	*0.980*	*0.090*	*0.883*
Victoria	*0.010*	*0.021*	*0.004*	*0.843*	***0.003***	***0.904***	*0.015*	-
St.John's	*0.025*	*0.249*	*0.011*	*0.839*	***0.006***	***0.951***	*0.020*	*0.497*
Halifax	*0.100*	*0.080*	*0.043*	*0.832*	***0.029***	***0.920***	*0.052*	*0.751*
Barrie	*0.012*	-	*0.004*	*0.496*	***0.003***	***0.599***	*0.007*	-
Markham	*0.008*	-	*0.003*	*0.809*	***0.003***	***0.874***	*0.020*	-

NOTE: RESULTS IN BOLD ARE THE BEST PERFORMING MODELS FOR EACH UTILITY

ERROR! NOT A VALID BOOKMARK SELF-REFERENCE. shows the average rate of failure for each network and can be used for better evaluating the RMSE of each model. The average current rate of failure for the Barrie network is 0.011, with an RMSE score of 0.003. Thus, since average current failure is significantly lower in this network, the result can not be compared with Winnipeg, which experienced a higher rate of failure. Or, in Winnipeg, more historical

information could have been collected that led to a higher failure rate. In a nutshell, the results from regression analysis should be interpreted with the help of practitioners from the industry.

TABLE 5.6 – AVERAGE OF CURRENT RATE OF FAILURES FOR ALL UTILITIES

Utility	Average of Current Rate of Failure (Number of failures/ Meter)
Barrie	0.011
Calgary	0.013
Region of Durham	0.013
Halifax	0.023
Kitchener	1.01
Markham	0.009
Region of Waterloo	0.085
Saskatoon	0.07
St. John's	0.019
Vancouver	0.174
Victoria	0.012
Waterloo	0.023
Winnipeg	0.082

The most important features for predicting the current rate of failure are the length and number of previous failures. The number of previous failures was removed from the analysis after creating the dependent variable. However, length was kept in this part of the study. Keeping length in the analysis increased the accuracy of the model. Consequently, in most cities, length was found to be highly influential in the final results. However, no specific overfitting was found for the models.

Other attributes do not significantly affect the final result, indicating that length may not be used for creating the target attribute, or even it should have been removed.

5.4 COMPARE RESULTS TO PREVIOUS STUDIES

Given tables compare the result of previous studies and one sample from this project. As seen within the tables, different studies achieved various performances. This difference is primarily due to the different algorithms used for each project. For instance, earlier studies in 1982 used simple linear regression, which considers each explanatory variable independently within the calculation. The authors in these studies suggested that the impact of all variables on each other should be analyzed precisely, as they seem to have an underlying relationship. On the other hand, other studies employed more complicated algorithms that use intricate mathematical steps to decrease the error and find the most likely pattern within the datasets.

These models have proven to be more reliable, and they are also capable of providing the most accurate prediction, if possible. Furthermore, the other difference between the current study and the previous studies is the number of available attributes, which varies among different projects.

Moreover, in terms of performance, the result of this study is comparable with previous studies. However, it did not show the best performance.

For the probability of failure, on the other hand, this study achieved better results compared to some of the recent studies. F1-Score in this study for the city of Saskatoon is 85% as opposed to the lower performance achieved by the mentioned studies.

For the rate of failure also, the approach of creating the target (dependent variable) was unique to this study. Nevertheless, the result of the present study indicated better performance when compared to some of the older studies. This could be as a result of having sufficient historical information for some of the utilities. The more information is collected, the better pattern can an algorithm find within the dataset.

Seemingly, as time elapses and more advanced algorithms are introduced, the accuracy and performance of the models are also enhanced.

TABLE 5.7 – COMPARING THE RESULTS OF PREVIOUS STUDIES (AGE TO FIRST FAILURE)

Approach	**Authors**	**Accuracy**
Linear Regression	Clark et al. (1982)	$R^2 = 0.23$
Linear Regression	McMullen (1982)	$R^2 = 0.375$
Ridge Regression	Almheiri et al., (2020)	$R^2 = 0.9$
ANN	Almheiri et al., (2020)	$R^2 = 0.84$
EDT (Ensemble Decision Tree)	Almheiri et al., (2020)	$R^2 = 0.88$
ANN	Snider and McBean, (2018)	$R^2 = 0.76$
XGBOOST	Present Study	$R^2 = 0.67$

TABLE 5.8 – COMPARING THE RESULTS OF PREVIOUS STUDIES (RATE OF FAILURE)

Approach	Authors	Accuracy
Linear Regression	Yamijala et al. (2009)	$R^2 = 0.12$
Linear Regression	Asnaashari et al. (2009)	$R^2 = 0.52$ to 0.88
Linear Regression	Giraldo-González et al. (2020)	$R^2 = 0.693$
Poisson Regression	Giraldo-González and Rodríguez, (2020)	$R^2 = 0.923$
Poisson Regression	Asnaashari et al. (2009)	$R^2 = 0.71$ to 0.95
Linear Regression	Bubtiena et al. (2011)	$R^2 = 0.737$
Linear Regression	Asnaashari et al. (2013)	$R^2 = 0.75$
EPR	Giraldo-González and Rodríguez, (2020	$R^2 = 0.877$
Poisson Regression	Giraldo-González and Rodríguez, (2020	$R^2 = 0.923$
Linear Regression	Kettler & Goulter (1985)	$R^2 = 0.93$ and 0.88
Weibull, Poisson and Yule	Martins et al. (2013)	MAE = 0.127 to 0.245
XGBOOST	Present Study	$R^2 = 0.958$

TABLE 5.9 – COMPARING THE RESULT OF PREVIOUS STUDIES (PROBABILITY OF FAILURE)

Approach	Authors	Accuracy
Naiev Bayes	Giraldo-González & Rodríguez, (2020)	F1-Score = 25.66%
Artificial Neural Networks (ANN)	Giraldo-González & Rodríguez, (2020	F1-Score = 42.10%
SVM	Giraldo-González & Rodríguez, (2020	F1-Score = 66.67%
Gradient Boosting	Giraldo-González & Rodríguez, (2020	F1-Score = 72%
XGBOOST	Present Study	F1- Score = 85%

6.

CONCLUSIONS

6.1 SUMMARY AND CONCLUSIONS

Drinking water networks are buried underground and cannot be easily assessed. Therefore, it is not feasible to frequently assess the condition of entire networks. Predictive deterioration models provide a cost-effective solution. Applying physical models requires a comprehensive, detailed dataset which is usually unavailable. Nevertheless, many statistical models have been developed thus far in order to find a global and efficient method to better predict the condition and evaluate the remaining useful life of a pipe. Nonetheless, the lack of a uniform model is still an obstacle for applying these models in practice.

Accordingly, this thesis focused on applying the most advanced machine learning models to predict the probability of failure, age at first failure, and the current rate of failures for thirteen utilities across Canada in order to better assess their broad accuracy and applicability. Data was collected from thirteen cities, including Saskatoon, Vancouver, Region of Waterloo, Waterloo, Region of Durham, Winnipeg, Markham, Halifax, Calgary, St.John's, Kitchener, Barrie, and Victoria. Results for all cities were then compared to evaluate how different algorithms perform with changing the utility.

Foremost, it should be mentioned that almost 80% of the project duration was dedicated to data cleaning. This step is critical, as having a comprehensive and problem-free dataset plays a vital role in the modeling process and future prediction. In order to have consistent datasets, unique values were defined for all attributes. These attributes include but are not limited to material, soil type, failure type, failure cause, and lining material. After cleaning the datasets and replacing unique values, missing values were found and replaced based on the information provided in GIS files.

To predict the probability of failure, XGBOOST, random forest (RF), artificial neural networks (ANN), and logistic regression (LR) were applied. XGBOOST showed the best performance, according to accuracy and F1-Score. One of the other advantages of these classification models

is the ability to provide information about the influence of each attribute in final prediction and during the learning process. Based on the results of classification models, *material, lining status (lining material), age, and length* were found to be the most influential factors to the learning and prediction steps. In previous studies, the material was frequently found as an important factor that emphasizes the importance of this attribute in this project. Age also has always been a controversial factor in many studies. Depending on the predominant type of material within each network, the importance of age may vary. However, this attribute indicated that it could be used for future prediction. HGL and Roughness were also concluded to have an impact on the process. However, only a few cities provided this information. It should be said that the contribution of each material as the most important factor varies among all utilities. For instance, cast iron plays a more important role in one network than ductile iron, and in other networks, vice versa.

Moreover, in the case where sufficient information about the broken pipe was not available, logistic regression also indicated a high performance in detecting the broken pipes. However, it was not able to find the non-broken pipes precisely. Therefore, the overall results did not show performing satisfactorily.

As mentioned previously, classification models are able to provide the probability of belonging to each class, which in this case is the probability of failure. Moreover, it should be noted that the model's performance shows its power to predict the classes. Therefore, the probability of failure calculated by each model should be interpreted based on the model's performance.

Some of the utilities provided imbalanced datasets for classification models where the number of broken pipes was significantly fewer than non-broken pipes. In this case, Synthetic Minority Oversampling Technique (SMOTE) was applied to make the datasets more balanced. This would help models learn from all available classes within the dataset more efficiently by creating artificial data points. However, the results of this model indicated that even with oversampling technique, tackling challenges regarding imbalanced datasets could be a demanding and unfruitful process. This method worked better where the datasets were significantly imbalanced, such as the Region of Waterloo network. The ratio of broken to non-broken pipes

in this network was 2:100, which can be considered significantly imbalanced. Nevertheless, it should be noted that the oversampling or undersampling methods are justifiable for the domains that the imbalanced format of the dataset is atypical. However, having these types of datasets in the water domain is normal, and it is recommended to keep the dataset unchanged. Accordingly, a few studies mentioned that undersampling and oversampling techniques improve the performance of the mode. However, as mentioned before, data manipulation, which may lead to changing the pattern, is not recommended.

In order to evaluate how homogeneity impacts the model accuracy, cast iron pipes were selected as a uniform group. Applying the model to cast iron pipes revealed that making a homogenous group of pipes would enhance the model performance. However, it is critical that both classes have sufficient recorded information for an efficient and accurate classification model. For instance, for cast iron pipes, the available information is desirably enough to have a good model. On the other hand, for PVC pipes and HDPE pipes, there is not enough information for broken pipes. This is why making a uniform group depends on different criteria that should be taken into account. Otherwise, some materials may cause a decline in the model`s performance due to a lack of data, as seen in this study. The results at the end of the appendices compare the performance of different models for various materials.

More importantly, based on the explanatory data analysis, *cast iron and ductile iron* pipes seem to be experiencing a significantly higher rate of deterioration. Therefore, further analysis should focus on these pipes. Asbestos cement pipes also showed a high rate of failure in some utilities. However, the contribution of this material to the failure among all networks is not comparable to that of cast iron, ductile iron, and PVC pipes.

Interestingly, failures related to PVC pipes have been recorded mainly during the early stage of their life cycles. Therefore, many factors may be evaluated to manage these pipes better to prevent future early failures. Furthermore, it should be mentioned that many failures pertained

to joint and fitting failures. However, these failures were excluded from this study since the primary focus of this thesis was to assess the natural deterioration of water pipes.

Furthermore, one global model was created, for cast iron pipes, based on the available information for all utilities. The output of this model was then tested on other cities, and a significant enhancement was noticed, especially for utilities that are suffering from a lack of adequate historical information. Seemingly, creating one model with a significant number of data points would be a feasible approach. However, many criteria should be checked in order to ensure the reliability of this Global model, such as controlling bias-variance trade-off. For example, for the Region of Waterloo, the accuracy of the model was considerably low. Nonetheless, using the Global model for the prediction enhanced the performance and indicated a higher accuracy and F1-Score for this network. It should be noted that length, diameter, and age were selected as common attributes among all networks for creating the global model. Age with almost 80% contribution was found to be the most important factor during the learning process.

Finally, the results of the classification models with the output of probability of failure (classification models are capable of predicting classes. However, it is possible to extract the probability of belonging to each class) were combined with GIS files. This gives a visualized map including the pipes, showing the probability of failure for each individual pipe. With this map, practitioners would be able to detect the most critical areas of the city and prioritize maintenance steps based on the criticality of a specific pipeline.

Regression models were also applied to the datasets to predict age at first failure. ElasicNet, random forest, artificial neural networks, and XGBOOST were utilized and compared. Unfortunately, the results of these models do not show satisfactory performance. Nonetheless, random forest indicated a better performance in this step, with an R-Squared of 0.67 for the Region of Waterloo, which indicated the best performance among all utilities. Moreover, creating a homogenous group of pipe (cast iron) did not significantly enhance the model`s performance. In addition, material, length, and diameter were found to be the most influential

factors in this step. Finally, anode status was found to be relatively important in the utilities that provided this attribute.

Finally, regression models were applied to the datasets in order to predict the current rate of failures. As previously mentioned, this target variable was created based on the length of pipes and the previous number of failures. ElasicNet, random forest, artificial neural networks, and XGBOOST were utilized and compared in this part. In most cases, the accuracy of the models was considerably high, except for ElasticNet regression, which did not perform well compared to others. XGBOOST, like the classification part, indicated the best performance with a relatively high R-Squared score. It should be mentioned that this step should be considered a preliminary study for predicting the current rate of failure with this method. Although the previous number of failures was removed from the study, length was kept. Keeping length made this attribute highly influential in predicting the current rate of failure, and without considering length, the accuracy of the models would drop significantly.

Overall, XGBOOST provided better results for classification and regression models. Due to its complex learning process, this model can detect the most intricate patterns within the datasets. XGBOOST is the developed and more regularized version of the gradient boosting method (Swamynathan, 2019). This method, specifically, employs gradient descent to build sequential decision trees that decline residuals (Snider and McBean, 2018). This algorithm is reported to have less susceptibility to noises and outliers and a shorter time for the training process (Snider and McBean, 2020a). This could be the primary reason for making this algorithm outperform other models. Nevertheless, the ANN algorithm indicated that it could be a serious contender for the XGBOOST model. More information about this algorithm can be found in chapter 2 of this project. In addition, for both classification and regression models, attributes that indicated colinearity within datasets were removed.

Finally and most importantly, tools utilized for conducting this project are found significantly practical. For example, python, as one of one the most growing tools for coding, has proven to be applicable for machine learning in this domain. In addition, power BI and GIS tools also provided relatively satisfactory outcomes in this project and may be applied for future studies.

6.1 LIMITATIONS

This study included a variety of input variables for both classification and regression models—however, the data varied by the utility. Thus, the same type of model applied to different cities cannot be easily compared since the inputs are different. Furthermore, data on both broken and non-broken pipes were required for classification models. However, some information such as soil type was only collected for broken pipes in certain utilities. This also limited the ability to apply this data.

Some of the broken pipes within this study were replaced during the maintenance process. However, information regarding these pipes was not available. Thus, these pipes were removed from the analysis. Future utility data collection should maintain historical pipe records to improve statistical models further.

The step of hyperparameters tuning was time-consuming since many parameters should have been tuned and evaluated for different models among all utilities. Nonetheless, all efforts have been made to find the best parameters for each model. However, this tuning process could have been done more profoundly and rigorously, especially for ANN and XGBOOST models. For example, finding the most optimum number of hidden layers and perceptrons is the most challenging process of the ANN model and could be further tuned.

5.2 FUTURE RECOMMENDATIONS

Based on the results and limitations of the present study, the following future areas of research and improvement to practice are recommended.

Data Collection

- Include soil-related attributes such as soil corrosivity, soil resistivity, moisture, and pH. These attributes should be collected for all available pipes, not only broken pipes. In most cases, some information was only collected for broken pipes in this study, which is not enough to achieve a comprehensive approach.

- Records of replaced pipes should not be removed from the datasets. This historical data should be maintained by utilities as it can further improve the accuracy of statistical deterioration models.

- Weather information can be collected and integrated into historical failure records in order to find out how different weather conditions may affect the deterioration process of water mains. As previously mentioned in the literature review, different seasons have various impacts on the pipe deterioration process. This emphasized the integration of weather information into historical records.

Modeling

- Use dimensionality reduction approaches to eliminate attributes that are not highly influential in the prediction. This method decreases the time required for training and increases the flexibility of the model to be tuned more efficiently.

- The developed models provided satisfactory predictions, which were the result of model tuning and validation for each utility. To avoid overfitting and underfitting, future models should focus on these steps.

- The influence of length on the current rate of failure should also be further investigated since this was found to be an important factor.

REFERENCES

Achim, D., Ghotb, F., & McManus, K. J. (2007). Prediction of Water Pipe Asset Life Using Neural Networks. *Journal of Infrastructure Systems*, *13*(1), 26–30. https://doi.org/10.1061/(ASCE)1076-0342(2007)13:1(26)

Ahn, J. C., Lee, S. W., Lee, G. S., & Koo, J. Y. (2005). *Predicting water pipe breaks using neural network*. 14.

Al-Barqawi, H., & Zayed, T. (2006). Condition Rating Model for Underground Infrastructure Sustainable Water Mains. *Journal of Performance of Constructed Facilities*, *20*(2), 126–135. https://doi.org/10.1061/(ASCE)0887-3828(2006)20:2(126)

Al-Barqawi, H., & Zayed, T. (2008). Infrastructure Management: Integrated AHP/ANN Model to Evaluate Municipal Water Mains' Performance. *Journal of Infrastructure Systems*, *14*(4), 305–318. https://doi.org/10.1061/(ASCE)1076-0342(2008)14:4(305)

Almheiri, Z., Meguid, M., & Zayed, T. (2020). Intelligent Approaches for Predicting Failure of Water Mains. *Journal of Pipeline Systems Engineering and Practice*, *11*(4), 04020044. https://doi.org/10.1061/(ASCE)PS.1949-1204.0000485

Andreou S. A. (1986). *Predictive models for pipe break failures and their implications on maintenance planning strategies for deteriorating water distribution systems*.

Andreou, S. A., Marks, D. H., & Clark, R. M. (1987a). A new methodology for modelling break failure patterns in deteriorating water distribution systems: Applications. *Advances in Water Resources*, *10*(1), 11–20. https://doi.org/10.1016/0309-1708(87)90003-0

Andreou, S. A., Marks, D. H., & Clark, R. M. (1987b). A new methodology for modelling break failure patterns in deteriorating water distribution systems: Theory. *Advances in Water Resources*, *10*(1), 2–10. https://doi.org/10.1016/0309-1708(87)90002-9

ASCE. (2009). *ASCE report card*. American Society of Civil Engineers. http://ascelibrary.org/doi/book/10.1061/9780784410370

ASCE. (2017). *ASCE drinking water report card*.

Asnaashari, A., McBean, E. A., Gharabaghi, B., & Tutt, D. (2013). Forecasting watermain failure using artificial neural network modelling. *Canadian Water Resources Journal*, *38*(1), 24–33. https://doi.org/10.1080/07011784.2013.774153

Asnaashari, A., & Shahrour, I. (2007). Analysis of Water Mains Failure Frequencies: Artificial Neural Networks Versus Poisson Regression, Case Study — Sanandaj-Iran. *Volume 14: Safety Engineering, Risk Analysis and Reliability Methods*, 195–202. https://doi.org/10.1115/IMECE2007-43402

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

Barton, N. A., Farewell, Timothy Stephen, Hallett, Stephen Henry, & Acland, Timothy Francis. (2019). Improving pipe failure predictions: Factors affecting pipe failure in drinking water networks. *Water Research*, 16.

Berkson, J. (1944). Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association*, *39*(227), 357–365. https://doi.org/10.1080/01621459.1944.10500699

Bhattarai, J. (2013). Study on the Corrosive Nature of Soil Towards the Buried-Structures. *Scientific World*, *11*(11), 43–47. https://doi.org/10.3126/sw.v11i11.8551

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). *Random Forest*.

Breiman, L., Friedman, Stone, & Olshen. (1984). *Classification And Regression Trees*. 369.

Bruaset, S. & Sægrov, S. (2018). An Analysis of the Potential Impact of Climate Change on the Structural Reliability of Drinking Water Pipes in Cold Climate Regions. *Water*, *10*(4), 411. https://doi.org/10.3390/w10040411

Canada infrastructure report card. (2016).

Canada infrastructure report card. (2019). 56.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Clark et al. (1982). *Water Distribution Systems: A Spatial and Cost Evaluation*.

Cox, D. R. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, *20*(2), 215–242.

D. R. Cox. (1972). *Regression Models and Life-Tables*.

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: Gradient boosting with categorical features support. *ArXiv:1810.11363 [Cs, Stat]*. http://arxiv.org/abs/1810.11363

Doyle, G., Seica, M. V., & Grabinsky, M. W. (2003). The role of soil in the external corrosion of cast iron water mains in Toronto, Canada. *Canadian Geotechnical Journal*, *40*(2), 225–236. https://doi.org/10.1139/t02-106

Fahmy, M., & Moselhi, O. (2009). Forecasting the Remaining Useful Life of Cast Iron Water Mains. *Journal of Performance of Constructed Facilities*, *23*(4), 269–275. https://doi.org/10.1061/(ASCE)0887-3828(2009)23:4(269)

Fernandez, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, *61*, 863–905. https://doi.org/10.1613/jair.1.11192

Freund, Y., & Schapire, R. E. (1996). *Experiments with a New Boosting Algorithm*. 16.

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

Freund, Y., & Schapire, R. E. (1999). *A Short Introduction to Boosting*. 14.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Fuchs-Hanusch, D., Friedl, F., Scheucher, R., Kogseder, B., & Muschalla, D. (2013). Effect of seasonal climatic variance on water main failure frequencies in moderate climate regions. *Water Supply*, *13*(2), 435–446. https://doi.org/10.2166/ws.2013.033

Giraldo-González, M. M., & Rodríguez, J. P. (2020). Comparison of Statistical and Machine Learning Models for Pipe Failure Modeling in Water Distribution Networks. *Water*, *12*(4), 1153. https://doi.org/10.3390/w12041153

Gould, S. J. F., Boulaire, F. A., Burn, S., Zhao, X. L., & Kodikara, J. K. (2011). Seasonal factors influencing the failure of buried water reticulation pipes. *Water Science and Technology*, *63*(11), 2692–2699. https://doi.org/10.2166/wst.2011.507

Goulter, I. C., & Kazemi, A. (1988). Spatial and temporal groupings of water main pipe breakage in Winnipeg. *Canadian Journal of Civil Engineering*, *15*(1), 91–97. https://doi.org/10.1139/l88-010

Goulter, I., Davidson, J., & Jacobs, P. (1993). Predicting Water-Main Breakage Rates. *Journal of Water Resources Planning and Management*, *119*(4), 419–436. https://doi.org/10.1061/(ASCE)0733-9496(1993)119:4(419)

Gummow, R. A., & Eng, P. (2004). *Control of External Corrosion On Iron and Steel Watermains*. 20.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Hu, Y., & Hubble, D. W. (2007). Factors contributing to the failure of asbestos cement water mains. *Canadian Journal of Civil Engineering*, *34*(5), 608–621. https://doi.org/10.1139/l06-162

Iličić, K. (2009). The analysis of influential factors on the frequency of pipeline failures. *Water Supply*, *9*(6), 689–698. https://doi.org/10.2166/ws.2009.779

InfraGuide. (2003). *Deterioration And Inspection Of Water Distribution Systems*. 34.

Jafar, R., Shahrour, I., & Juran, I. (2010). Application of Artificial Neural Networks (ANN) to model the failure of urban water mains. *Mathematical and Computer Modelling*, *51*(9–10), 1170–1180. https://doi.org/10.1016/j.mcm.2009.12.033

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Jarosz, B. (2021). Poisson Distribution: A Model for Estimating Households by Household Size. *Population Research and Policy Review*, *40*(2), 149–162. https://doi.org/10.1007/s11113-020-09575-x

Kabir, G., Tesfamariam, S., Loeppky, J., & Sadiq, R. (2016). Predicting water main failures: A Bayesian model updating approach. *Knowledge-Based Systems*, *110*, 144–156. https://doi.org/10.1016/j.knosys.2016.07.024

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

Kerwin, S., Garcia de Soto, B., Adey, B., Sampatakaki, K., & Heller, H. (2020). Combining recorded failures and expert opinion in the development of ANN pipe failure prediction models. *Sustainable and Resilient Infrastructure*, 1–23. https://doi.org/10.1080/23789689.2020.1787033

Kettler, A. J., & Goulter, I. C. (1985). An analysis of pipe breakage in urban water distribution networks. *Canadian Journal of Civil Engineering*, *12*(2), 286–293. https://doi.org/10.1139/l85-030

Kimutai, E., Betrie, G., Brander, R., Sadiq, R., & Tesfamariam, S. (2015). Comparison of Statistical Models for Predicting Pipe Failures: Illustrative Example with the City of Calgary Water Main Failure. *Journal of Pipeline Systems Engineering and Practice*, *6*(4), 04015005. https://doi.org/10.1061/(ASCE)PS.1949-1204.0000196

Kleinbaum, D. G., & Klein, M. (2010). *Logistic Regression*. Springer New York. https://doi.org/10.1007/978-1-4419-1742-3

Kleiner, & Rajani. (2001). Comprehensive review of structural deterioration of water mains: Statistical models. *Urban Water*, *3*(3), 131–150. https://doi.org/10.1016/S1462-0758(01)00033-4

Kleiner, & Rajani. (2002). Forecasting Variations and Trends in Water-Main Breaks. *Journal of Infrastructure Systems*, *8*(4), 122–131. https://doi.org/10.1061/(ASCE)1076-0342(2002)8:4(122)

Laucelli, D., Rajani, B., Kleiner, Y., & Giustolisi, O. (2014). Study on relationships between climate-related covariates and pipe bursts using evolutionary-based modelling. *Journal of Hydroinformatics*, *16*(4), 743–757. https://doi.org/10.2166/hydro.2013.082

Le Gat, Y., & Eisenbeis, P. (2000). Using maintenance records to forecast failures in water networks. *Urban Water*, *2*(3), 173–181. https://doi.org/10.1016/S1462-0758(00)00057-1

Lei, J. and Saegrov, S. (1998). *Statistical approach for describing failures and lifetimes of water mains*. 9.

Lisa A. Jeffrey. (1985). *Predicting urban water distribution maintenance strategies, A case study of New Haven, Connecticut*.

Mailhot, A., Pelletier, G., Noël, J.-F., & Villeneuve, J.-P. (2000). Modeling the evolution of the structural state of water pipe networks with brief recorded pipe break histories: Methodology and application. *Water Resources Research*, *36*(10), 3053–3062. https://doi.org/10.1029/2000WR900185

Martínez-Codina, Á., Castillo, M., González-Zeas, D., & Garrote, L. (2016). Pressure as a predictor of occurrence of pipe breaks in water distribution networks. *Urban Water Journal*, *13*(7), 676–686. https://doi.org/10.1080/1573062X.2015.1024687

Mazumder, R. K., Salman, A. M., Li, Y., & Yu, X. (2019). Reliability Analysis of Water Distribution Systems Using Physical Probabilistic Pipe Failure Method. *Journal of Water Resources Planning and Management*, *145*(2), 04018097. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001034

Mena, L., & Gonzalez. (2006). *Machine Learning for Imbalanced Datasets: Application in Medical Diagnostic*. 6.

Mordak, J. & Wheeler, J. (1988). *Deterioration of Asbestos Cement Water Mains*. Water research council.

Moselhi, O., & Shehab-Eldeen, T. (2000). Classification of Defects in Sewer Pipes Using Neural Networks. *Journal of Infrastructure Systems*, *6*(3), 97–104. https://doi.org/10.1061/(ASCE)1076-0342(2000)6:3(97)

Motiee, H., & Ghasemnejad, S. (2019). Prediction of pipe failure rate in Tehran water distribution networks by applying regression models. *Water Supply*, *19*(3), 695–702. https://doi.org/10.2166/ws.2018.137

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Najafi, M. & Kulandaivel, G. (2005). *Pipeline Condition Prediction Using Neural Network Models*.

Nishiyama, M., & Filion, Y. (2013). Review of statistical water main break prediction models. *Canadian Journal of Civil Engineering*, *40*(10), 972–979. https://doi.org/10.1139/cjce-2012-0424

Nishiyama, M., & Filion, Y. (2014). Forecasting breaks in cast iron water mains in the city of Kingston with an artificial neural network model. *Canadian Journal of Civil Engineering*, *41*(10), 918–923. https://doi.org/10.1139/cjce-2014-0114

Park, H.-A. (2013). An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain. *Journal of Korean Academy of Nursing*, *43*(2), 154. https://doi.org/10.4040/jkan.2013.43.2.154

Park, Jun, H., Agbenowosi, N., Kim, B. J., & Lim, K. (2011). The Proportional Hazards Modeling of Water Main Failure Data Incorporating the Time-dependent Effects of Covariates. *Water Resources Management*, *25*(1), 1–19. https://doi.org/10.1007/s11269-010-9684-y

Park, Kim, J. W., Newland, A., Kim, B. J., & Jun, H. D. (2008). Survival Analysis of Water Distribution Pipe Failure Data Using the Proportional Hazards Model. *World Environmental and Water Resources Congress 2008*, 1–10. https://doi.org/10.1061/40976(316)500

Park, S. (2004). Identifying the hazard characteristics of pipes in water distribution systems by using the proportional hazards model: 1. Theory. *KSCE Journal of Civil Engineering*, *8*(6), 663–668. https://doi.org/10.1007/BF02823557

Philip, B. E., & Aljassmi, H. (2020). *The Relevance of Water Pipe Deterioration Prediction Models: A review*. *9*(02), 5.

Pritchard, O. G., & Hallett, D. S. H. (2013). *Soil Corrosivity in the UK – Impacts on Critical Infrastructure*. 55.

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

Rajeev, P., Kodikara, J., Robert, D., Zeman, P., & Rajani, B. (n.d.). *FACTORS CONTRIBUTING TO LARGE DIAMETER WATER PIPE FAILURE AS EVIDENT FROM FAILURE INSPECTION*. 11.

Rebala, G., Ravi, A., & Churiwala, S. (2019). *An Introduction to Machine Learning*. Springer International Publishing. https://doi.org/10.1007/978-3-030-15729-6

Rezaei, H., Ryan, B., & Stoianov, I. (2015). Pipe Failure Analysis and Impact of Dynamic Hydraulic Conditions in Water Supply Networks. *Procedia Engineering*, *119*, 253–262. https://doi.org/10.1016/j.proeng.2015.08.883

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

Røstum, J. (2000). *STATISTICAL MODELLING OF PIPE FAILURES IN WATER NETWORKS*. 132.

Ruchti, G. F. (2017). *Water Pipeline Condition Assessment*. American Society of Civil Engineers. https://doi.org/10.1061/9780784414750

Rui Wang, Weishan Dong, Yu Wang, Ke Tang, & Xin Yao. (2013). Pipe failure prediction: A data mining method. *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, 1208–1218. https://doi.org/10.1109/ICDE.2013.6544910

Sadler, J. M., Goodall, J. L., Morsy, M. M., & Spencer, K. (2018). Modeling urban coastal flood severity from crowd-sourced flood reports using Poisson regression and Random Forest. *Journal of Hydrology*, *559*, 43–55. https://doi.org/10.1016/j.jhydrol.2018.01.044

Saeed Mirza. (2007). *Danger ahead: The coming collapse of Canada's municipal infrastructure*. Federation of Canadian Municipalities. https://www.deslibris.ca/ID/250220

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

Shamir and Howard. (1979). An Analytic Approach to Scheduling Pipe Replacement. *Journal - American Water Works Association*, *71*(5), 248–258. https://doi.org/10.1002/j.1551-8833.1979.tb04345.x

Sharma, S., Sharma, S., Scholar, U., & Athaiya, A. (2020). *ACTIVATION FUNCTIONS IN NEURAL NETWORKS*. *4*(12), 7.

Shi, F., Liu, Y., Liu, Z., & Li, E. (2018). Prediction of pipe performance with stacking ensemble learning based approaches. *Journal of Intelligent & Fuzzy Systems*, *34*(6), 3845–3855. https://doi.org/10.3233/JIFS-169556

Shirzad, A., & Safari, M. J. S. (2019). Pipe failure rate prediction in water distribution networks using multivariate adaptive regression splines and random forest techniques. *Urban Water Journal*, *16*(9), 653–661. https://doi.org/10.1080/1573062X.2020.1713384

Snider, B., & McBean, E. A. (2018). *IMPROVING TIME-TO-FAILURE PREDICTIONS FOR WATER DISTRIBUTION SYSTEMS USING GRADIENT BOOSTING ALGORITHM*. 8.

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

Snider, B., & McBean, E. A. (2020b). Improving Urban Water Security through Pipe-Break Prediction Models: Machine Learning or Survival Analysis. *Journal of Environmental Engineering*, *146*(3), 04019129. https://doi.org/10.1061/(ASCE)EE.1943-7870.0001657

St. Clair, A. M., & Sinha, S. (2012). State-of-the-technology review on water pipe condition, deterioration and failure rate prediction models! *Urban Water Journal*, *9*(2), 85–112. https://doi.org/10.1080/1573062X.2011.644566

St. Clair, A. M., & Sinha, S. (2014). Development of a Standard Data Structure for Predicting the Remaining Physical Life and Consequence of Failure of Water Pipes. *Journal of Performance of Constructed Facilities*, *28*(1), 191–203. https://doi.org/10.1061/(ASCE)CF.1943-5509.0000384

Steven Folkman. (2012). *Water Main Break Rates In the USA and Canada A Comprehensive Study*. 28.

Steven Folkman. (2018). *Water Main Break Rates In the USA and Canada: A Comprehensive Study*. 48.

Swamynathan, M. (2019). *Mastering Machine Learning with Python in Six Steps: A Practical Implementation Guide to Predictive Data Analytics Using Python*. Apress. https://doi.org/10.1007/978-1-4842-4947-5

Syachrani, S., Jeong, H. S. "David," & Chung, C. S. (2013). Decision Tree–Based Deterioration Model for Buried Wastewater Pipelines. *Journal of Performance of Constructed Facilities*, *27*(5), 633–645. https://doi.org/10.1061/(ASCE)CF.1943-5509.0000349

Tabesh, M., Soltani, J., Farmani, R., & Savic, D. (2009). Assessing pipe failure rate and mechanical reliability of water distribution networks using data-driven modeling. *Journal of Hydroinformatics*, *11*(1), 1–17. https://doi.org/10.2166/hydro.2009.008

Task Committee on Water Pipeline Condition Assessment. (2017). *Water Pipeline Condition Assessment* (G. F. Ruchti, Ed.). American Society of Civil Engineers. https://doi.org/10.1061/9780784414750

Tavakoli, R. (2018). *REMAINING USEFUL LIFE PREDICTION OF WATER PIPES USING ARTIFICIAL NEURAL NETWORK AND ADAPTIVE NEURO FUZZY INFERENCE SYSTEM MODELS*. 169.

Tran, D. H., Ng, A. W. M., & Perera, B. J. C. (2007). Neural networks deterioration models for serviceability condition of buried stormwater pipes. *Engineering Applications of Artificial Intelligence*, *20*(8), 1144–1151. https://doi.org/10.1016/j.engappai.2007.02.005

Tran, H. D., Perera, B. J. C., & Ng, A. W. M. (2009). Predicting Structural Deterioration Condition of Individual Storm-Water Pipes Using Probabilistic Neural Networks and Multiple Logistic Regression Models. *Journal of Water Resources Planning and Management*, *135*(6), 553–557. https://doi.org/10.1061/(ASCE)0733-9496(2009)135:6(553)

Vanrenterghem-Raven, A., Eisenbeis, P., Juran, I., & Christodoulou, S. (2003). Statistical Modeling of the Structural Degradation of an Urban Water Distribution System: Case Study of New York City. *World Water & Environmental Resources Congress 2003*, 1–10. https://doi.org/10.1061/40685(2003)41

Verdhan, V. (2020). *Supervised Learning with Python: Concepts and Practical Implementation Using Python*. Apress. https://doi.org/10.1007/978-1-4842-6156-9

Vladeanu, G. J., & Koo, D. D. (2015). A Comparison Study of Water Pipe Failure Prediction Models Using Weibull Distribution and Binary Logistic Regression. *Pipelines 2015*, 1590–1601. https://doi.org/10.1061/9780784479360.146

Walski, T. M., & Pelliccia, A. (1982). Economic analysis of water main breaks. *Journal (American Water Works Association)*, *74*(3,), 140–147.

Wang, W., Robert, D., Zhou, A., & Li, C.-Q. (2018). Factors Affecting Corrosion of Buried Cast Iron Pipes. *Journal of Materials in Civil Engineering*, *30*(11), 04018272. https://doi.org/10.1061/(ASCE)MT.1943-5533.0002461

Wang, Y., Moselhi, O., & Zayed, T. (2009). Study of the Suitability of Existing Deterioration Models for Water Mains. *Journal of Performance of Constructed Facilities*, *23*(1), 40–46. https://doi.org/10.1061/(ASCE)0887-3828(2009)23:1(40)

Wasim, M., Shoaib, S., Mubarak, N. M., Inamuddin, & Asiri, A. M. (2018). Factors influencing corrosion of metal pipes in soils. *Environmental Chemistry Letters*, *16*(3), 861–879. https://doi.org/10.1007/s10311-018-0731-x

Wengström. (1993). *Drinking Water Pipe Breakage Records-A Tool for Evaluating Pipe and System Reliability*.

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

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

Wols, B. A., van Daal, K., & van Thienen, P. (2014). Effects of Climate Change on Drinking Water Distribution Network Integrity: Predicting Pipe Failure Resulting from Differential Soil Settlement. *Procedia Engineering*, *70*, 1726–1734. https://doi.org/10.1016/j.proeng.2014.02.190

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

Wols, B. A., & van Thienen, P. (2014b). Impact of weather conditions on pipe failure: A statistical analysis. *Journal of Water Supply: Research and Technology-Aqua*, *63*(3), 212–223. https://doi.org/10.2166/aqua.2013.088

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

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

Zhang, D., Qian, L., Mao, B., Huang, C., Huang, B., & Si, Y. (2018). A Data-Driven Design for Fault Detection of Wind Turbines Using Random Forests and XGboost. *IEEE Access*, *6*, 21020–21031. https://doi.org/10.1109/ACCESS.2018.2818678

Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, *58*, 308–324. https://doi.org/10.1016/j.trc.2015.02.019

APPENDIX A – DATA DESCRIPTION (ALL CITIES IN DETAIL)

Saskatoon – Saskatchewan

Inventory File

Saskatoon water pipeline, as one of the largest networks, is located in Saskatchewan province in Canada. This city has a population of around 246 thousand with a land area of about 228 square kilometers. The raw inventory file of Saskatoon includes 35,630 pipe segments with a variety of collected features, such as Diameter, Material, Joint Type, Installation Year, Replaced Year, Replaced Status, Ownership, Length, Lining Year, and Lining Material. The given table provides more details about the ranges and values of the attributes mentioned above (TABLE 0.1).

TABLE 0.1 – AVAILABLE ATTRIBUTES WITHIN SASKATOON INVENTORY DATASET

Attribute	Unit	Range/Values
Diameter	mm	25 - 1350
Material	Text	PVC, AC, CI, ST, PE, HDPE, DI, CON, PVCF
Join Type	Text	Rubber Universal Lead Mechanical Grooved Threaded Welded
Installation Year	Year	1906 - 2019
Replaced Year	Year	1928 - 2018
Replaced Status	Binary	Yes/No
Ownership	Binary	Yes/No
Length	m	0.2 - 1581

Lining Year	Year	2004 - 2017
Lining Material	Text	CIP, HDPE, and PE

All above information included within the datasets is extracted from the network GIS files (shapefiles). Figure 0.1 illustrates the GIS map of the Saskatoon water main network.

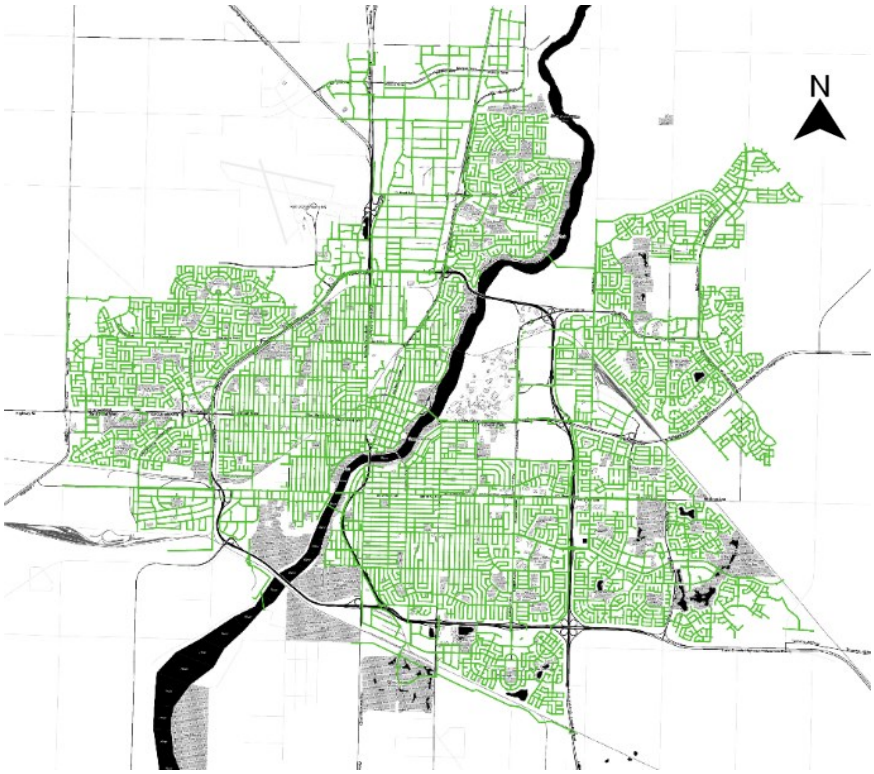


FIGURE 0.1 – SASKATOON WATER DISTRIBUTION NETWORK (GIS FILE PROVIDED BY CITY OF SASKATOON)

The city of Saskatoon owns 1,193 kilometers of water mains, which consists of PVC (46%), Asbestos Cement (31%), Cast Iron (16%), Steel (3%), and other (Copper, PE, HDPE, PVCF, Ductile Iron - 3%). The given figure shows the percentage of each material based upon joint type and percentage of total length (Figure 0.2). As can be seen, Rubber is the most popular joint type for PVC and other materials. However, threaded is reported to have been the most frequent type for asbestos cement pipe. For Cast Iron, there is no specific pattern, although lead and Universal joints seem to be the methods that have been used more frequently for this material.

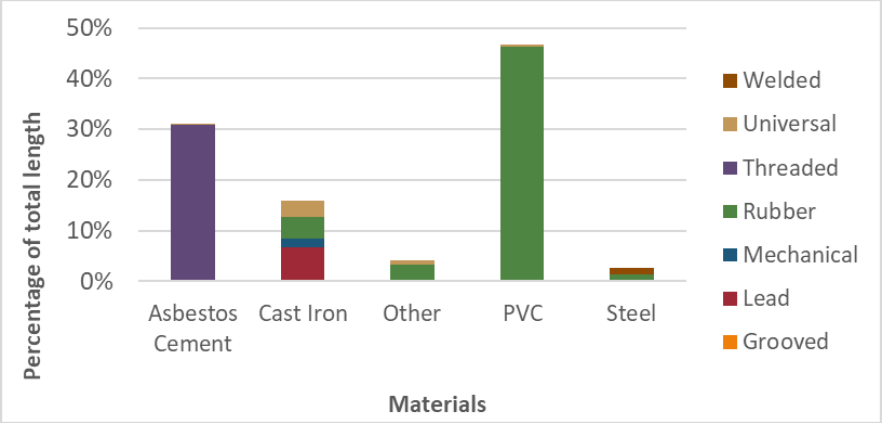


FIGURE 0.2 – PERCENTAGE OF EACH MATERIAL BASED ON LENGTH AND JOIN TYPE WITHIN SASKATOON NETWORK (INVENTORY FILE)

As previously mentioned, pipes were installed from 1906 to 2019 in the Saskatoon network. The frequency of different pipe materials installed in different time intervals can be seen in Figure 0.3. As can be seen from the stack chart, from 1900 to 1960, Cast Iron was a predominant material installed in Saskatoon. It should be noted that from 1940 to 1960, Asbestos Cement was as much popular as Cast Iron. Afterward, AC experienced a significant increase for 20 years, between 1960 and 1980. However, after introducing PVC pipes, the city of Saskatoon seems to have decided to focus on installing this material, as it has demonstrated a substantial surge from 1980 to 2019, with being a primary material to be installed.

Another important attribute provided by the Saskatoon city is diameter, and it ranges from 25 mm to 1350 mm. As can be seen from the chart (Figure 0.4), most pipes range from 150 to 300, typically related to the water distribution network. Pipes with 150 mm are the most frequent size used by Saskatoon utility.

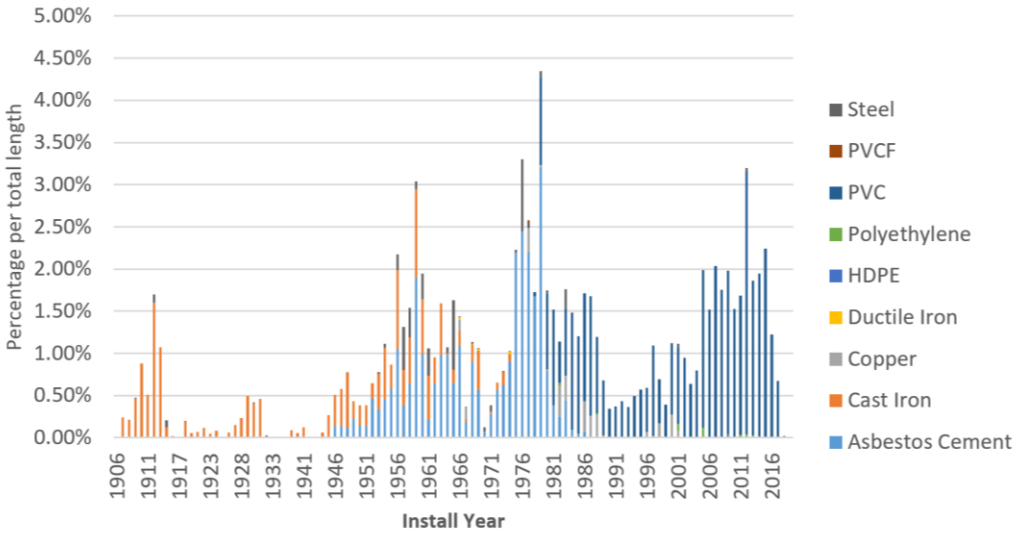


FIGURE 0.3 – PERCENTAGE OF EACH MATERIAL PER TOTAL LENGTH BASED ON INSTALLATION YEAR (SASKATOON – INVENTORY)

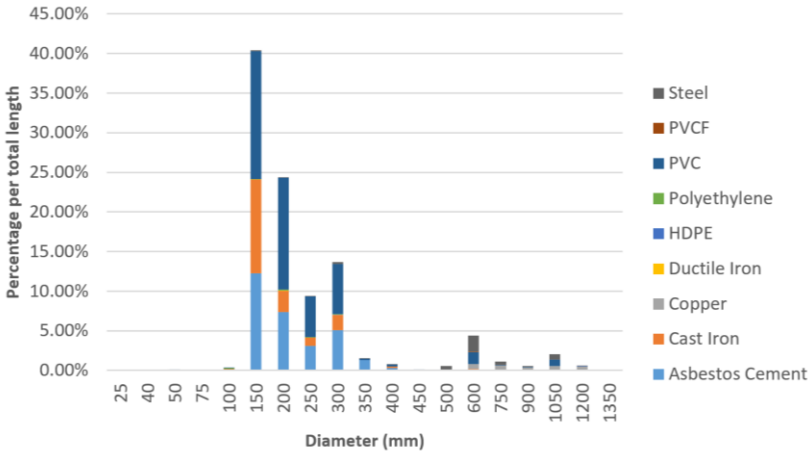


FIGURE 0.4 – PERCENTAGE OF EACH MATERIAL BASED ON SIZE AND TOTAL LENGTH (SASKATOON)

In addition, some of the pipes were either lined during their life cycles, or even installed lined since the beginning, and some pipes remained unlined during their entire lives. More than 90% of pipe reported to have remained unlined, and only around 10% have lining protection. CIP (cured in place), HDPE (high density polyethylene), and PE are the lining types that have been used in Saskatoon.

Break File

Historical failure records were provided as a break file by utilities. After the data cleaning process, the Saskatoon break file comprises 7,083 data points, and the given figure indicates the percentage of each material that experienced failure within the network (Figure 0.5). According to historical records, cast iron pipes account for just over 46% of the total failures, followed by AC and PVC pipes, with 41.66% and 7.5%, respectively. Steel material makes up only 4% of all failure records, and other materials account for only 1% of instances. It should be mentioned that a significant proportion of failure records, in terms of material and diameter, has been extracted from the inventory GIS file due to having many missing values in the break file.

The provided failure records included material, diameter, failure date, and pipe depth. More attributes are added from the inventory file to calculate the failure rate, age at failure, and the probability of failure. It is worth mentioning that the number of failures for each unique pipe ID is from 1 to 21 failures within the Saskatoon network. For example, there exists a Cast Iron pipe that has experienced 21 failures during its life cycle. Figure 0.5 demonstrates the percentage of each material based on different types of failures. As it can be seen from the chart, hole and circumferential failures are the most frequent types among others. For instance, the hole is the most frequent type of failure for PVC, cast iron, and asbestos cement. It is worth mentioning that a significant proportion of pipes do not include the type of failure. Therefore, these unknown failures were named “Other”.

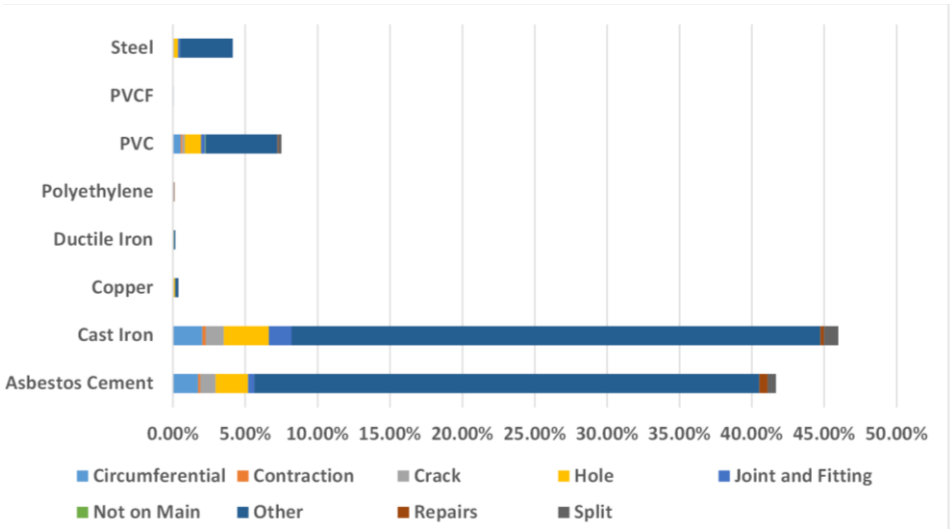


FIGURE 0.5 - PERCENTAGE OF EACH MATERIAL IN THE NETWORK AND THEIR CORRESPONDING FAILURES (BREAK FILE – SASKATOON)

Failures are provided from 1958 to 2019 for the Saskatoon network. Figure 0.6 shows the percentage of total failures in each year for different types of materials.

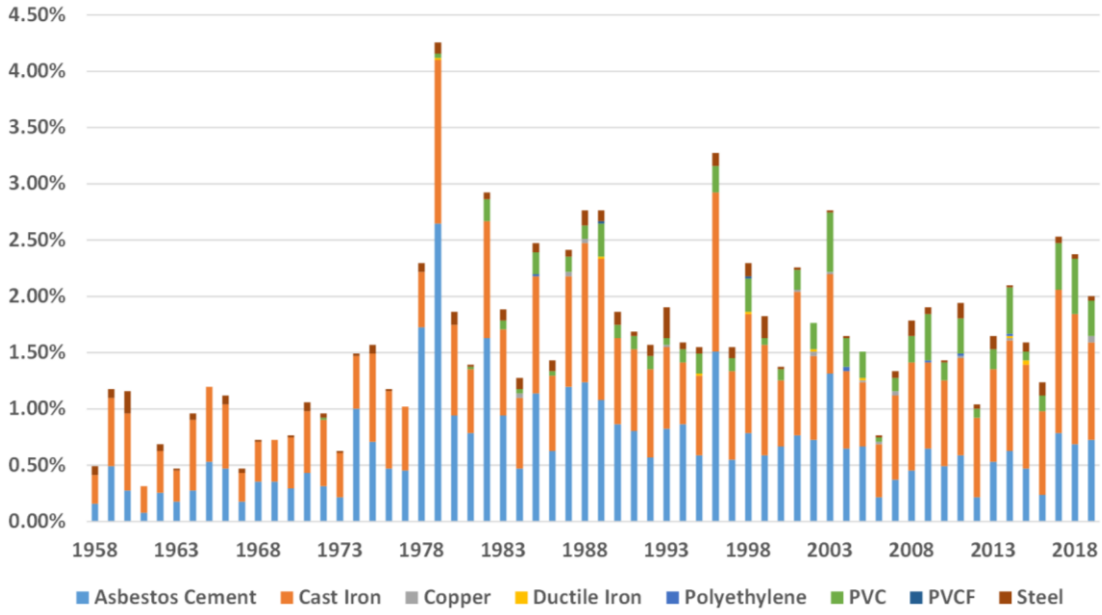


FIGURE 0.6 – PERCENTAGE OF CONTRIBUTION OF EACH YEAR IN FAILURE RECORD BASED ON THE MATERIALS (SASKATOON – BREAK FILE)

Winnipeg – Manitoba

Inventory File

Winnipeg, situated in Manitoba province, is another robust data set analyzed in this research. The population of this city is registered as approximately 705 thousand, based on the 2016 Census. The land area of this city is around 464 square kilometers. Furthermore, the base inventory file of Winnipeg consists of 114,824 segments, and various variables are provided such as material, diameter, joint type, status, status date, ownership, length, install year, and coating material. The provided table demonstrates the range and categories for the mentioned attributes in more detail (TABLE 0.2).

TABLE 0.20.2 - AVAILABLE ATTRIBUTES WITHIN WINNIPEG INVENTORY DATASET

Attribute	Unit	Range/Values
Diameter	mm	19 - 1050
Material	Text	PVC, AC, CI, ST, PE, HDPE, DI, CO, CON, PCCP, PB
Join Type	Text	Rubber Universal Lead Mechanical Grooved Threaded Welded Bell and Spigot Bell Flange Flared end Gasket Socket
Installation Year	Year	1882 - 2020
Ownership	Binary	Yes/No
Length	m	0.01 – 996.59
Status	Text	Active/InActive

Coating Material	Text	Asbestos, Concrete, FRC, polyethylene, Styrofoam, Urecon, and Y-jacket

As previously mentioned, the GIS version of the inventory file was utilized to create a robust dataset where missing values are available. The given map has been extracted from the GIS file, which shows the complexity of the Winnipeg network (Figure 0.7).

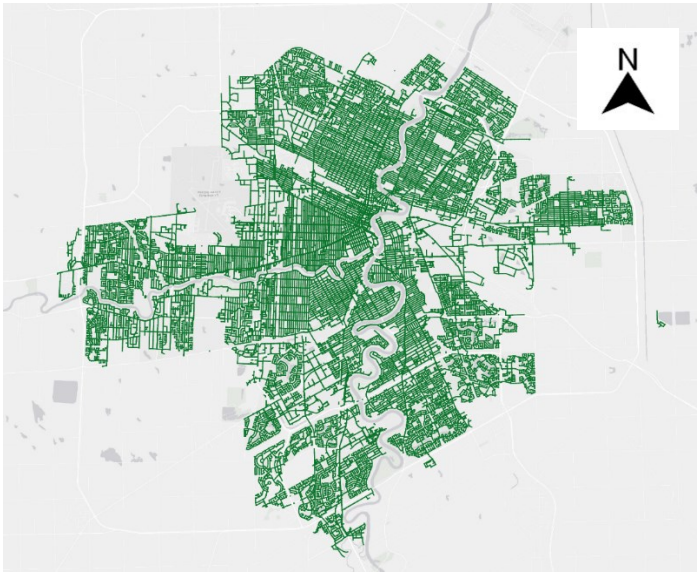


FIGURE 0.7 - WINNIPEG WATER DISTRIBUTION NETWORK (GIS FILE PROVIDED BY CITY OF WINNIPEG)

Winnipeg water networks consist of 3,117 kilometers of pipes, among which 48.77% are made from PVC. AC and CI pipes account for 23.73% and 26.08%, respectively, and the remainder of almost 2% belong to other materials.

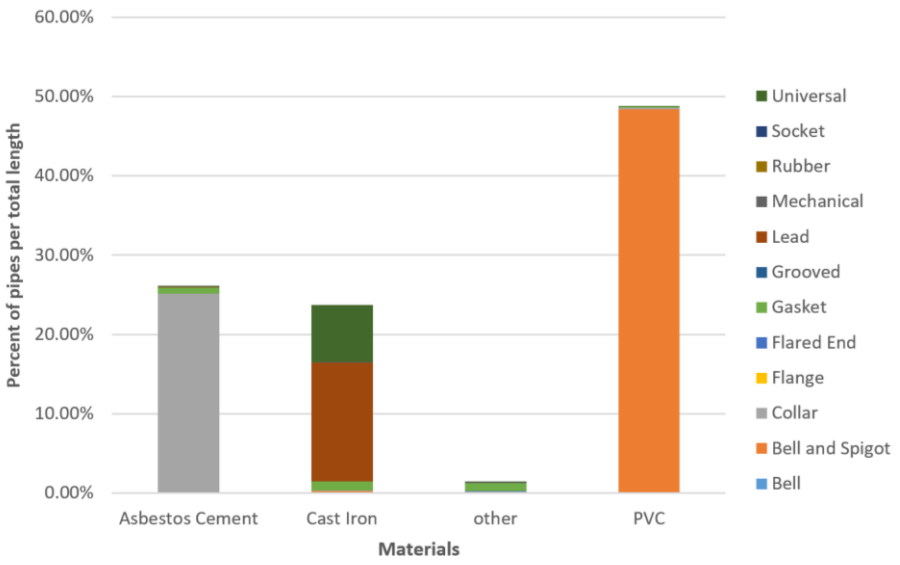


FIGURE 0.8 – DIFFERENT TYPES OF MATERIALS WITHIN WINNIPEG INVENTORY BASED ON TOTAL LENGTH AND JOINT TYPES

Provided information revealed that pipes were installed from 1882 to 2020 in the Winnipeg network. Installation of different materials experienced the same trend as in the Saskatoon network. Cast Iron pipes seem to have been the most frequent type of materials from 1882 to 1960, peaking during the 1950s. AC pipes also underwent a dramatic surge from 1946 to 1985. From 1986, however, PVC became the predominant material within this network (Figure 0.9).

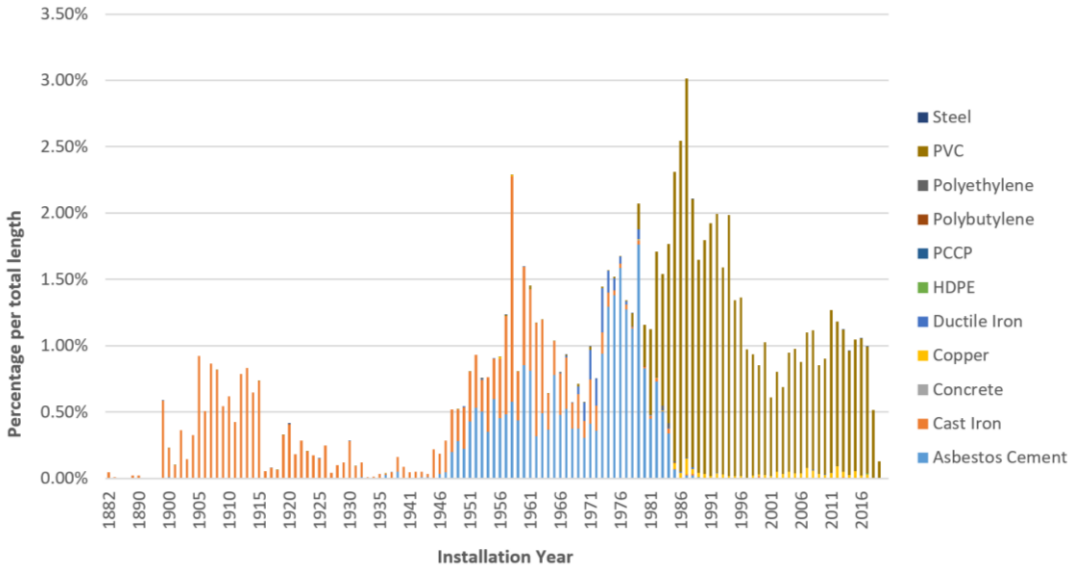


FIGURE 0.9 - PERCENTAGE OF EACH MATERIAL PER TOTAL LENGTH BASED ON INSTALLATION YEAR (WINNIPEG – INVENTORY)

In terms of diameter, the Winnipeg network comprises ranges from 19 to 1050 mm. The bar chart below shows that most pipes within the network fall in diameter between 150 and 300

mm. Other sizes account for a small proportion of the entire network (Figure 0.10). It should be noted that almost all pipes within the network are uncoated.

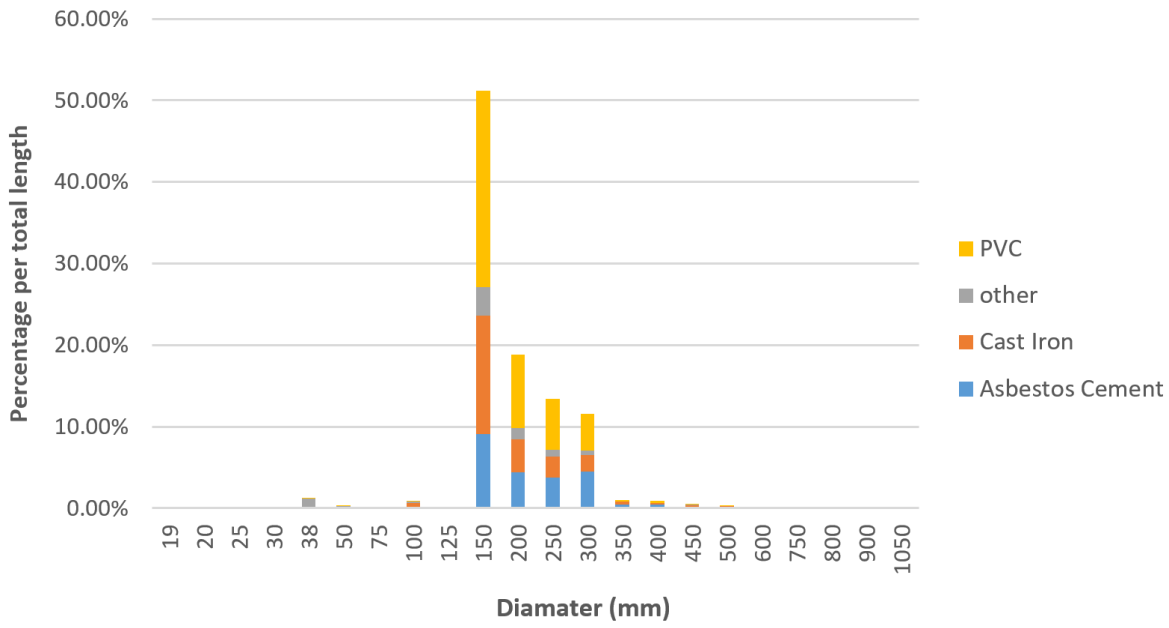


FIGURE 0.10 - PERCENTAGE OF EACH MATERIAL BASED ON SIZE AND TOTAL LENGTH (WINNIPEG)

Break File

The failure dataset provided by Winnipeg city consists of 26,631 pipes. Among this amount of records, almost 69% of failures are related to CI pipes. AC pipes also have a contribution of around 15.40% to the total amount of breaks. PVC and Ductile Iron make up 11.75% and 3.72% of historical failures, respectively. Several attributes are provided with break records, such as failure date, failure type, soil type, and land use. However, soil type and land use have been removed from the analysis since they include a significant amount of missing values. Holes seem to have been the most frequent type of failure within the Winnipeg network. Cast Iron and PVC also experienced a considerable rate of failures related to joint and fittings. Circumferential crack, however, is the predominant type of failure for AC pipes (Figure 0.11). There are many missing values for the type of failures within the Winnipeg network. Thus, these missing values were named “Other.” The following bar chart gives the percentage of each material in the network and its corresponding number of failures.

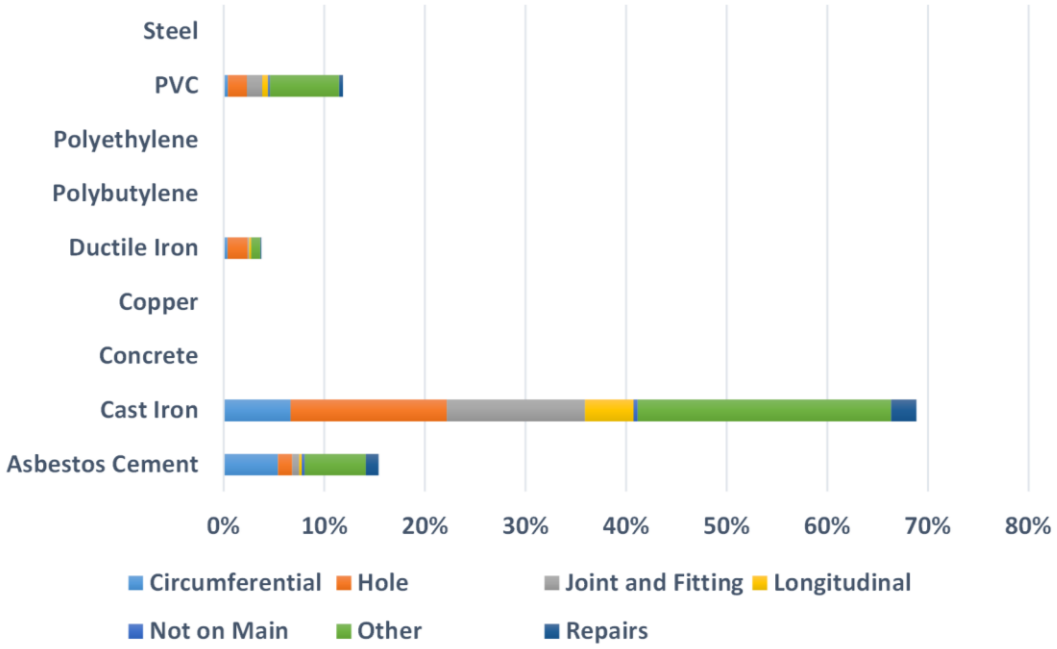


FIGURE 0.11 – PERCENTAGE OF EACH MATERIAL IN THE NETWORK AND THEIR CORRESPONDING FAILURES (BREAK FILE – WINNIPEG)

Failures records have been collected from 1919 to 2019 and provided with the dataset. From the chart below, it is clear that collecting information experienced a significant surge after 1989, with CI as a material that has undergone a significant number of failures. For AC pipes, there is a fluctuation in the number of failures recorded. Overall, it is clear that CI pipes are in an extreme deterioration process, and they should be prioritized towards any asset management practices. Figure 0.12 indicates the number of failures for different materials in different periods.

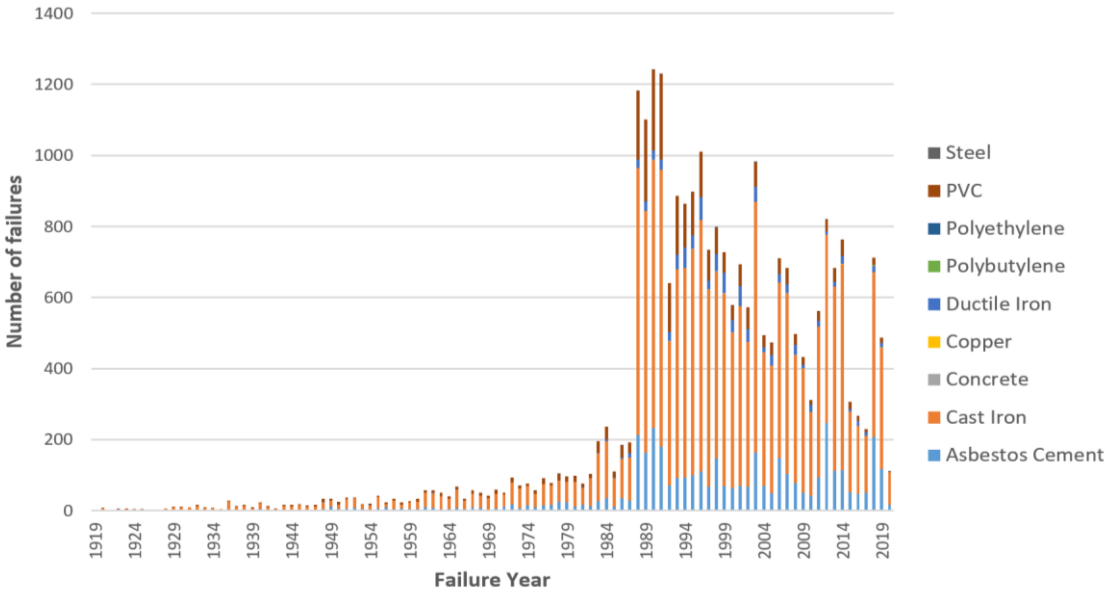


FIGURE 0.12 – NUMBER OF FAILURES FOR DIFFERENT MATERIALS IN DIFFERENT YEARS (WINNIPEG)

Kitchener – Ontario

Inventory File

Kitchener is another utility analyzed in this study and is located in Ontario province, Canada. Based on the 2016 census, the population of this city was reported to be around 233 thousand inhabitants, with a land area of about 137 square kilometers.

After cleaning, the inventory file of Kitchener included 14,561 pipes, including different input variables such as diameter, material, installation year, status, length, lining status, lining year, and lining material. The given table lists all attributes along with the range of provided values (TABLE 0.3).

TABLE 0.3~~0.3~~ - AVAILABLE ATTRIBUTES WITHIN KITCHENER INVENTORY DATASET

Attribute	Unit	Range/Values
Diameter	mm	25 - 1200
Material	Text	AC, CI, CON, CO, DI, HDPE, PVC, PVCB, PVCO, ST

Installation Year	Year	1887 - 2018
Status	Binary	Active, Inactive
Length	m	0.01 – 83.46
Lining Status	Binary	Yes, No
Lining Year	Year	1977 - 2014
Lining Material	Text	CM, EPOXY, Unlined

Figure 0.13 shows the percentage of each material based on the total length within the Kitchener network. The graph indicates that ductile iron pipe accounts for 37.21% of the entire network, followed by PVC, with a 31.43% contribution. Cast iron is another frequently used pipe within this network with almost 23.13% contribution to the total length.

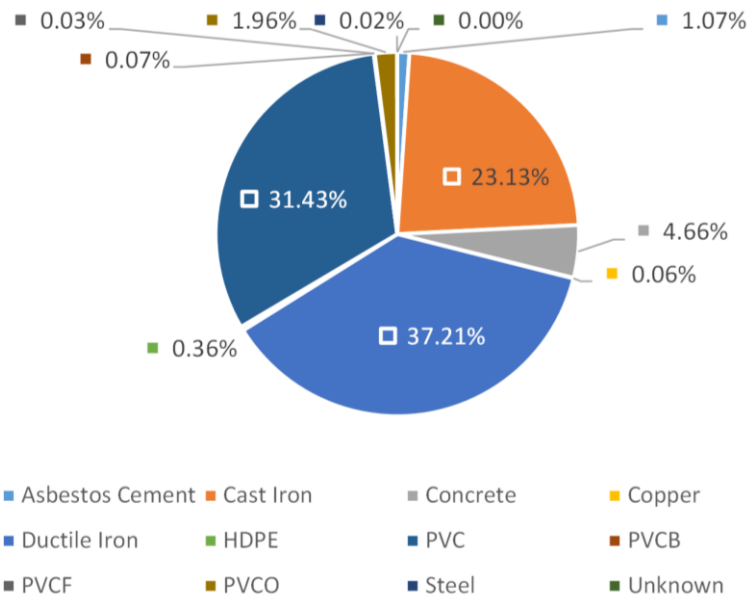


FIGURE 0.13 – DIFFERENT TYPES OF MATERIALS WITHIN KITCHENER INVENTORY BASED ON THE TOTAL LENGTH

Collected information revealed that pipes were installed in this network from 1887 to 2018. From the beginning until the '60s, cast iron was the primary material within this network. However, ductile iron altered the trend as it became the primary material utilized in Kitchener

from the `60s to the early `90s. Moreover, PVC pipe seems to have been the predominant pipe within this network during the 20th century. Nonetheless, from 1987 to 1992, concrete pipe showed an increase in installation for this network (Figure 0.14).

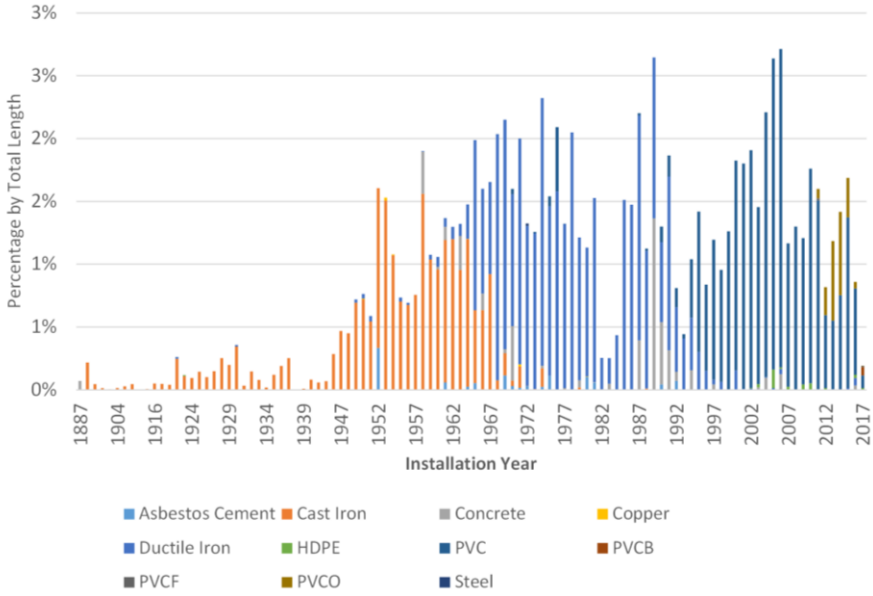


FIGURE 0.14 - PERCENTAGE OF EACH MATERIAL PER TOTAL LENGTH BASED ON INSTALLATION YEAR (KITCHENER - INVENTORY)

Regarding diameter, this network includes different sizes, ranging from 25 mm to 1200 mm. However, pipes with the size of 150 mm are the most frequent ones in the network with a 45% contribution to total length, for which cast iron, PVC, and ductile iron are equally distributed. 300 and 200 mm pipes are in the following position, accounting for 25% and 13% of the entire network, respectively (Figure 0.15).

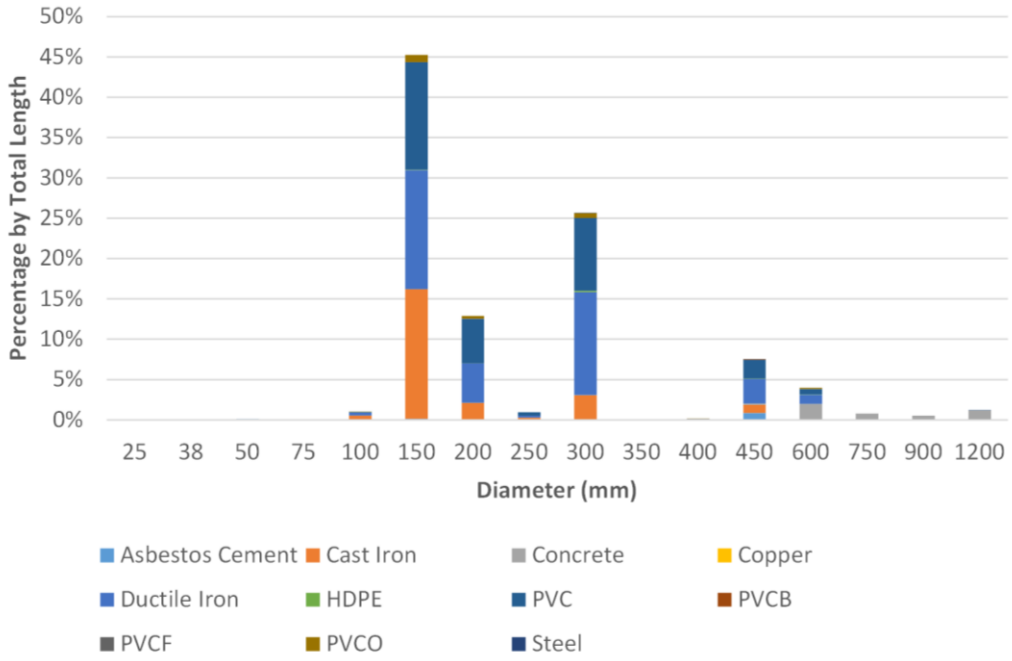


FIGURE 0.15 - PERCENTAGE OF EACH MATERIAL BASED ON SIZE AND TOTAL LENGTH (KITCHENER)

Break File

After cleaning, the excel file for the break records included 2,346 pipes in the Kitchener network. A significant proportion of failures are related to cast iron pipes, accounting for over 75% of total failures. Ductile iron is also with 25% contribution to total failures is the following material. The given bar chart indicates the proportion of each material in the network based on the number of failures and type of failures (Figure 0.16). As seen in the graph, most pipes do not have the type of failure or cause of failures, as shown by “Other” in the chart. Nonetheless, for cast iron and ductile iron, circumferential failures were the most frequent type of failure. Several attributes were provided along with this dataset, including status, pipe depth, anode status, failure date, failure type, failure cause, and soil type.

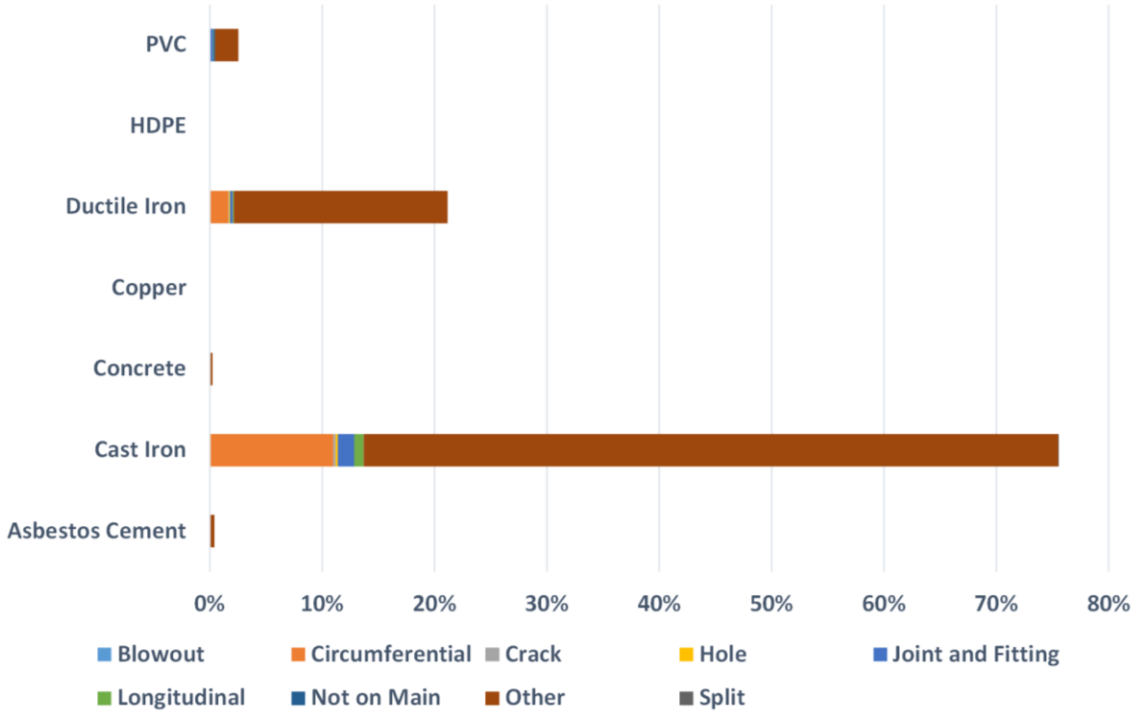


FIGURE 0.16 – PERCENTAGE OF EACH MATERIAL IN THE NETWORK AND THEIR CORRESPONDING FAILURES (BREAK FILE – KITCHENER)

The city of Kitchener provided failures records from 1985 to 2018, showing cast iron as a material that experienced a considerable number of failures compared to other materials. It seems that collecting information has undergone an increase from 1997 to date (Figure 0.17).

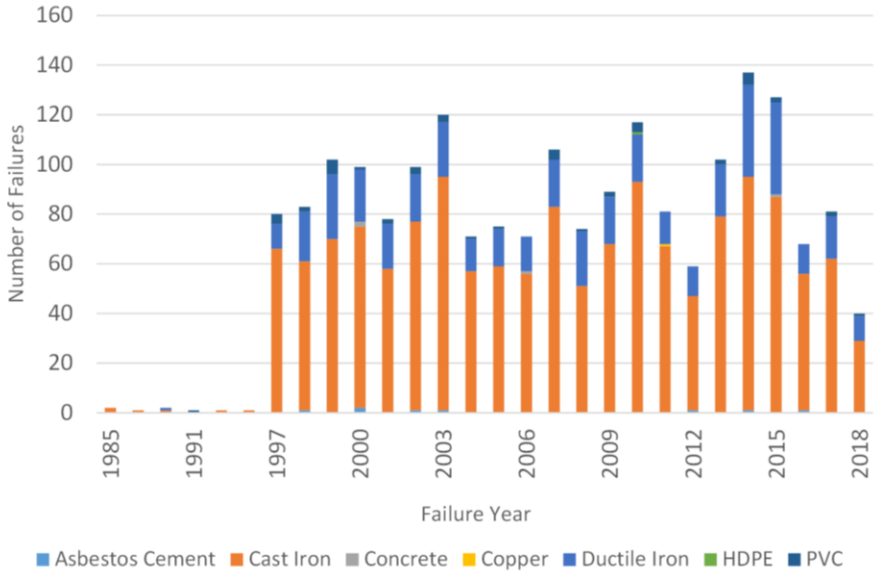


FIGURE 0.17 – NUMBER OF FAILURES FOR DIFFERENT MATERIALS IN DIFFERENT YEARS (KITCHENER)

Markham – Ontario

Inventory File

With a population of around 329 thousand and a land area of 212.35 square kilomteres, Markham is located in the Ontario province, Canada. After data cleaning and preparation, this utility included 10,802 pipe segments with a total length of approximately 1,261 kilometers. Additionally, as shown in the given table, diameter, material, roughness, installation year, ownership, length, lining and protection status, lining and protection year, and pipe depth are the attributes provided within the excel file (TABLE 0.4).

TABLE 0.4~~0.4~~ - AVAILABLE ATTRIBUTES WITHIN MARKHAM INVENTORY DATASET

Attribute	Unit	Range/Values
Diameter	mm	25 - 1800
Material	Text	AC, CI, CON, CO, CLPE, DI, HDPE, PE, PB, PVC, ST
Roughness	Text	39 - 187
Installation Year	Year	1938 - 2019
Ownership	Binary	Yes, No
Length	m	0.82 – 3779
Lining Status	Binary	Yes, No
Protection Status	Binary	Yes, No
Lining Year	Year	1996 - 2011
Protection Year	Year	1992 - 2015
Pipe Depth	m	0.85 – 4.20

PVC pipe is the most frequently installed material within the Markham network, accounting for 66.36% of the total length. This material is followed by ductile iron that makes up 18.58% of the entire network. Concrete and cast iron pipes are other types of materials that have been used, with 5.82% and 4.66% contribution, respectively. It is worth mentioning that the use of cast iron pipe in this network is not comparable to other networks in this study. The percentage of other types of materials is shown within the given pie chart (Figure 0.18).

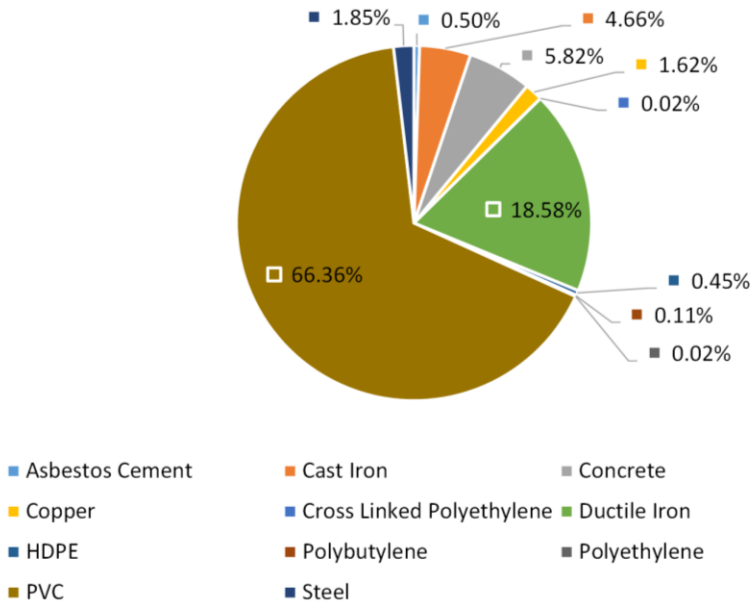


FIGURE 0.18 - DIFFERENT TYPES OF MATERIALS WITHIN MARKHAM INVENTORY BASED ON THE TOTAL LENGTH

From the beginning of the data collection until the mid-60s, a mixture of materials was used in the Markham network, including cast iron, steel, and ductile iron pipes. However, from 1965 until 1980, ductile iron was the most desirable material in this network. Since then, PVC seems to have become the predominant material used for water networks in the Markham utility (Figure 0.19).

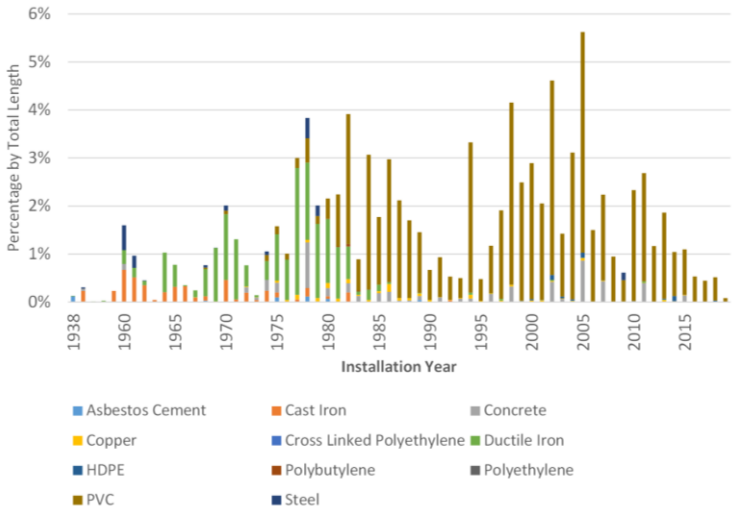


FIGURE 0.19 - PERCENTAGE OF EACH MATERIAL PER TOTAL LENGTH BASED ON INSTALLATION YEAR (MARKHAM - INVENTORY)

Figure 0.20 shows the distribution of pipes based on the size and their materials within the network. As shown in the graph, smaller pipes account for a significant portion of the entire network. For example, 150-mm pipes are the most frequently used in the network with almost 32% contribution, followed by 200 and 300 mm pipes. It should be noted that PVC, cast iron, and ductile iron have the highest contribution for smaller pipes. However, for larger pipes, concrete and steel pipes seem to be more prevalent within this network.

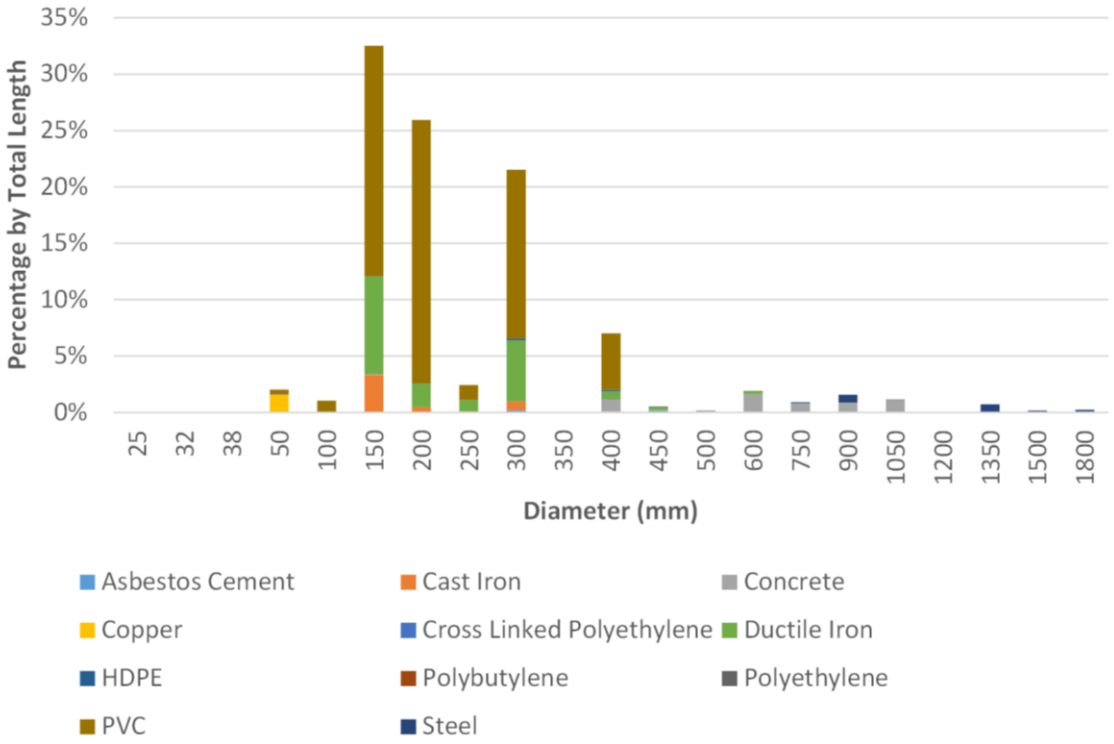


FIGURE 0.20 - PERCENTAGE OF EACH MATERIAL BASED ON SIZE AND TOTAL LENGTH (MARKHAM)

Break File

The final excel file for the Markham network included 2,926 failures record. Ductile iron accounts for more than 45% of the total failures in the network, followed by cast iron and PVC pipes with 39% and 13% contribution. For ductile iron, the hole is the most frequent type of failure. Also, some failures known as “Not on main” in this study relate to ductile iron, such as saddle failure. However, the pattern is different for cast iron, with circumferential and hole as the most frequent failures in this network. Saddle failure also has a notable contribution to cast iron failures. The given bar chart indicates the portion of each material in the network based on the total number of failures and type of failures (Figure 0.21).

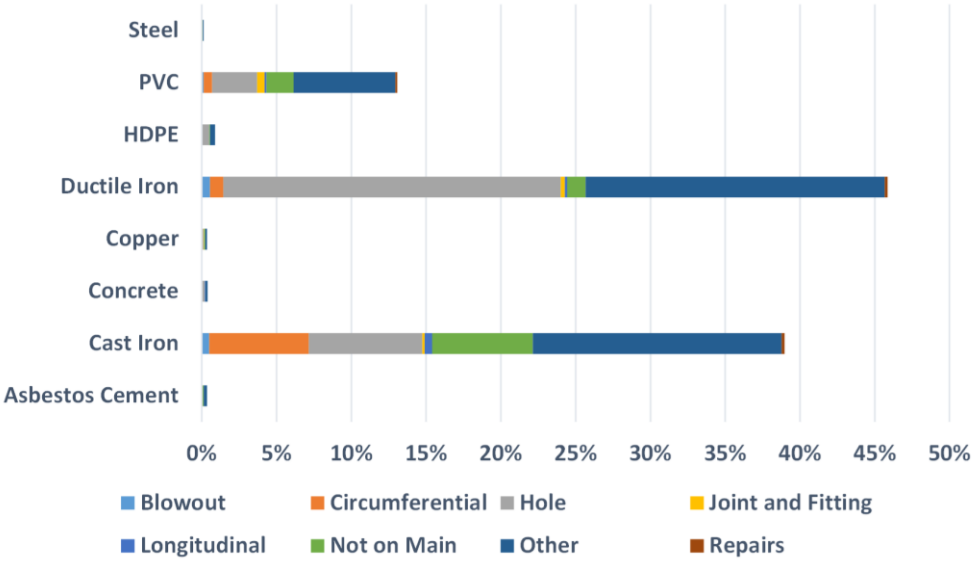


FIGURE 0.21 - PERCENTAGE OF EACH MATERIAL IN THE NETWORK AND THEIR CORRESPONDING FAILURES (BREAK FILE – MARKHAM)

Failures record provided from 1979 to 2019 in Markham utility. The given bar chart shows that the number of failures has undergone a downward trend, gradually declining from 1994 to 2018. As previously mentioned, cast iron and ductile iron experienced the most failures recorded within this utility.

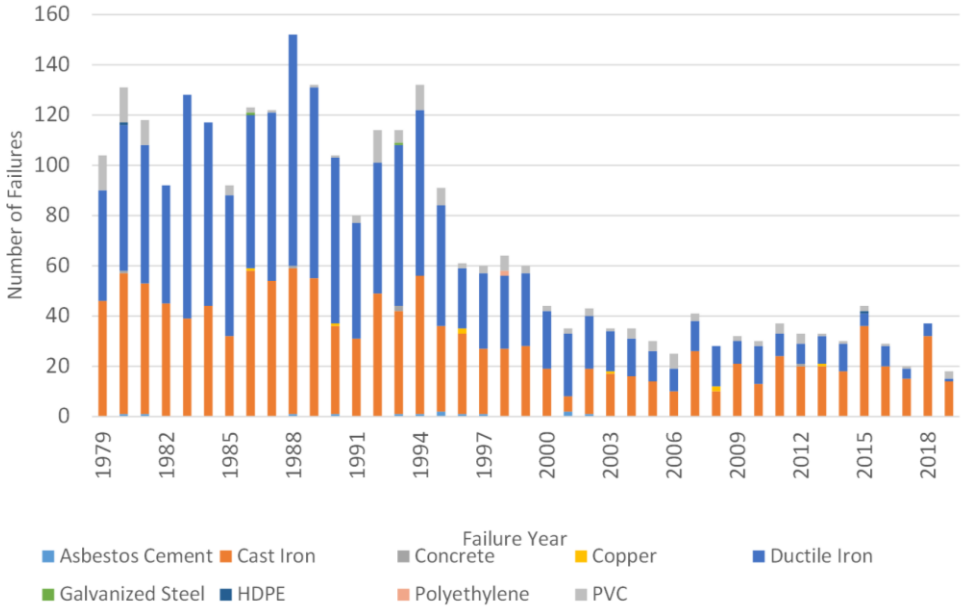


FIGURE 0.22 - NUMBER OF FAILURES FOR DIFFERENT MATERIALS IN DIFFERENT YEARS (MARKHAM)

Waterloo – Ontario

Inventory File

Waterloo is a city situated in Ontario province, Canada. This city has a population of around 113 thousand and a land area of around 64 square kilometers. Waterloo pipeline consists of 7,565 pipes with a total length of around 433 kilometers. Diameter, material, service type, installation year, ownership, length, lining status, lining material, and lining year are the attributes in the final cleaned dataset. The range of these features can be found in the given table (TABLE 0.5).

TABLE 0.5~~0.5~~ - AVAILABLE ATTRIBUTES WITHIN WATERLOO INVENTORY DATASET

Attribute	Unit	Range/Values
Diameter	mm	25 - 450
Material	Text	AC, CI, CON, CO, DI, HDPE, PE, PVC
Service Type	Text	Distribution, Transmission
Installation Year	Year	1850 - 2018
Ownership	Binary	Yes, No
Length	m	0.09 – 644
Lining Status	Binary	Yes, No
Lining Material	Text	CM, Epoxy, HDPE, Unlined
Lining Year	Year	1999 - 2013

Regarding the distribution of each material in the network, PVC is the most installed type. This material contributes to 54.29% of the entire network. Cast iron is the following material, which accounts for 29.68% of the total pipes. Finally, with 15.12% of the total length of the inventory file, ductile iron is another material used in the Waterloo network. Figure 0.23 shows the percentage of each material in the network based on the total length.

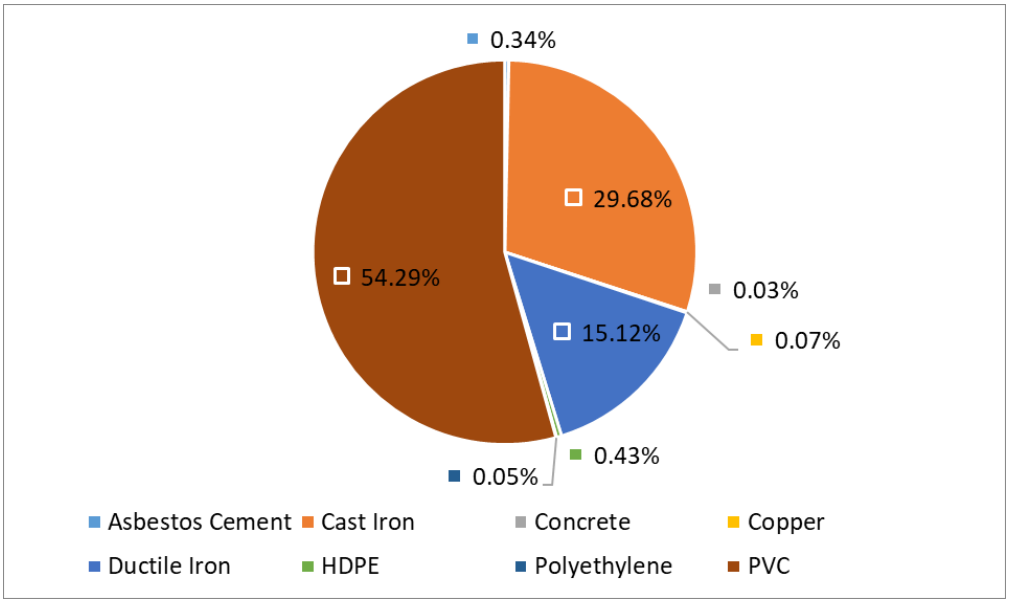


FIGURE 0.23 - DIFFERENT TYPES OF MATERIALS WITHIN WATERLOO INVENTORY BASED ON THE TOTAL LENGTH

From the beginning of the data collection in this network until the early 70s, cast iron was the most popular material. Then, from 1974 to 1984, ductile iron became the prevalent type of material in this utility. However, same as other utilities, from 1984 to date, PVC pipes have been the predominant type of material in the network. The distribution of different materials based on the installation year is provided in the given chart (Figure 0.24).

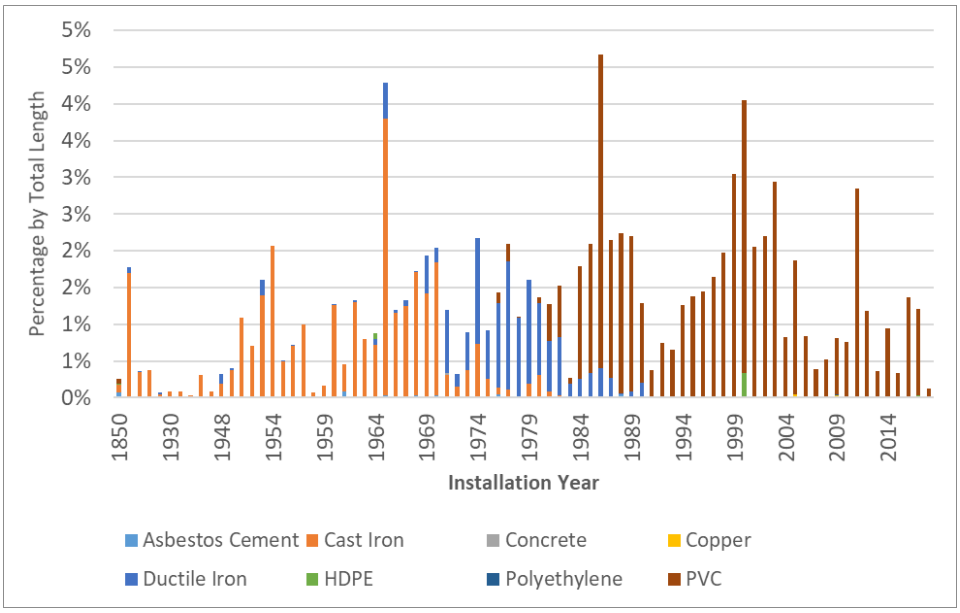


FIGURE 0.24 - PERCENTAGE OF EACH MATERIAL PER TOTAL LENGTH BASED ON INSTALLATION YEAR (WATERLOO - INVENTORY)

In terms of size, pipes with a diameter of 150 mm are the most frequent ones in the network. This size accounts for almost 45% of the entire utility. Diameter of 300 mm and 200 mm with 26% and 21% contribution, respectively, are the primarily other installed diameter in the networks. The majority of these pipes are related to the distribution network. The given bar chart shows the contribution of each diameter to the total length of pipes in the network (Figure 0.25).

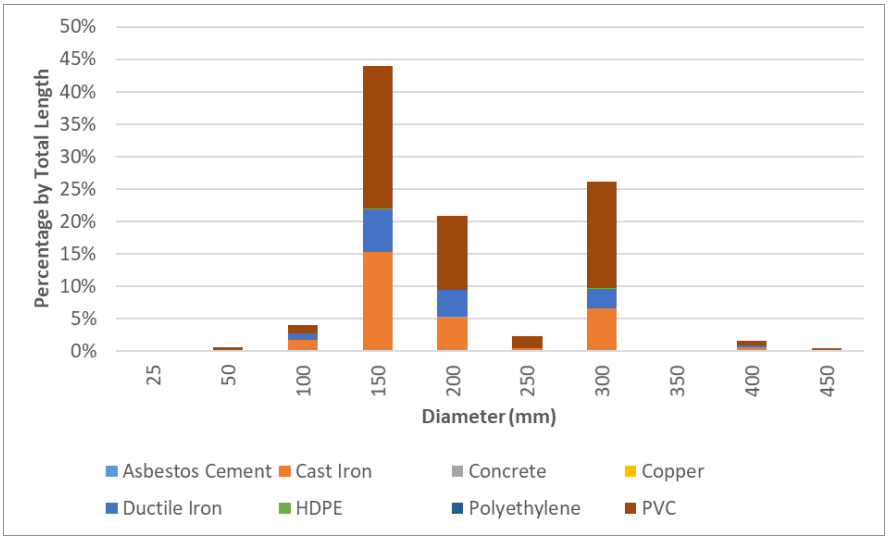
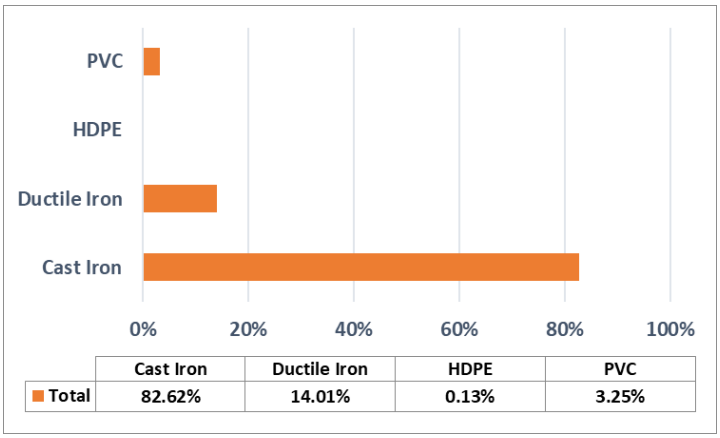


FIGURE 0.25 - PERCENTAGE OF EACH MATERIAL BASED ON SIZE AND TOTAL LENGTH (WATERLOO)

Break File

Waterloo network did not provide the failure type of the recorded failures. Therefore, the given graph is prepared only based on the total length of pipes within the network. The graph shows that cast iron pipes experienced the highest number of failures in this network, accounting for almost 83% of total failures (Figure 0.26). This material is followed by ductile iron, with around 14% of the total number of failures recorded in the network.



	Cast Iron	Ductile Iron	HDPE	PVC
Total	82.62%	14.01%	0.13%	3.25%

FIGURE 0.26 - PERCENTAGE OF EACH MATERIAL IN THE NETWORK (BREAK FILE – WATERLOO)

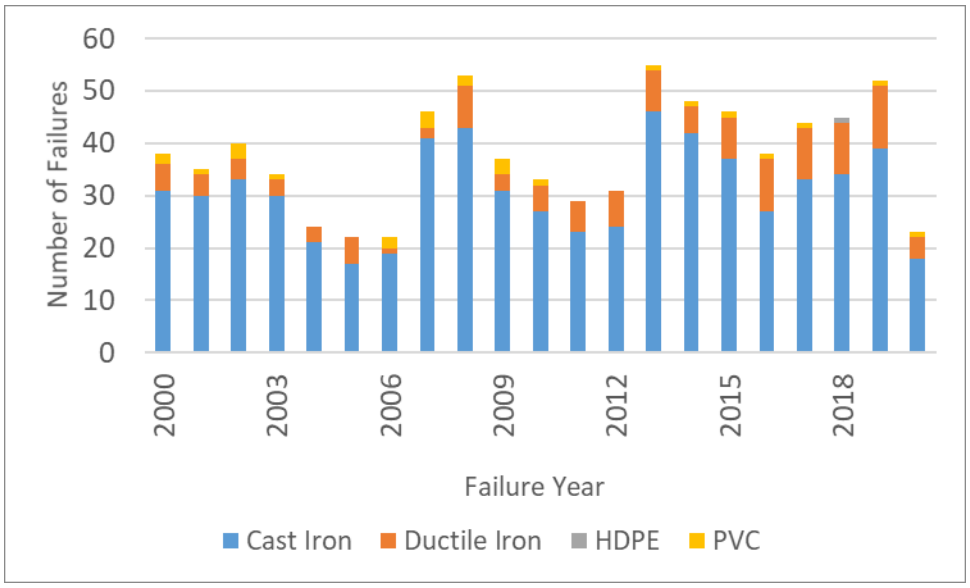


FIGURE 0.27 - NUMBER OF FAILURES FOR DIFFERENT MATERIALS IN DIFFERENT YEARS (WATERLOO)

Failures were provided from 2000 to 2020 for the Waterloo network. The given graph shows a fluctuation in the number of failures for different years. However, as previously discussed, cast iron underwent the highest failure rate in this network (Figure 0.27).

Region of Waterloo – Ontario

Inventory File

Region of Waterloo is a metropolitan area located in the Ontario province, Canada. With a population of around 523 thousand and a land area of above 1,300 square kilometers, this network consists of 5,139 pipes. These pipes have different attributes such as diameter, material, installation year, ownership, length, lining status, lining material, lining year, roughness, bedding type, surface type, soil type, and pipe depth. It should be noted the total length of this network is above 430 kilometers. The given table indicates more information regarding these attributes (TABLE 0.6).

TABLE 0.6~~0.6~~ - AVAILABLE ATTRIBUTES WITHIN REGION OF WATERLOO INVENTORY DATASET

Attribute	Unit	Range/Values
Diameter	mm	38 - 1200
Material	Text	AC, CI, CON, CO, DI, HDPE, PE, PVC, ST
Installation Year	Year	1850 - 2019
Ownership	Binary	Yes, No
Length	m	0.06 – 6977
Lining Status	Binary	Yes, No
Lining Material	Text	CM, Unlined
Lining Year	Year	1974 - 2004
Roughness	µ	110 - 150
Bedding Type	Text	Concrete, Granular
Surface Type	Text	Asphalt, Concrete, Exposed, Grass, Gravel, Road, Water
Soil Type	Text	Granular, Gravel
Pipe Depth	m	0.5 - 4

PVC and ductile iron are the most frequently used materials within this network, accounting for almost 43% and 25.57% of the toal length. Concrete is the following material with a 16.25

contribution. Cast iron and asbestos cement are equally distributed within the network, with almost 7% of the total length for each (Figure 0.28).

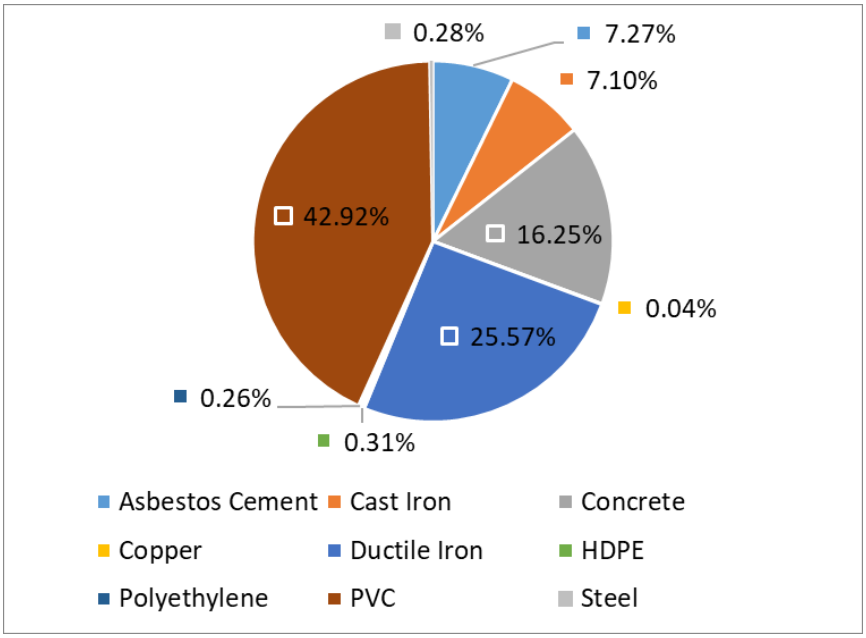


FIGURE 0.28 - DIFFERENT TYPES OF MATERIALS WITHIN REGION OF WATERLOO INVENTORY BASED ON THE TOTAL LENGTH

Similar to other networks, cast iron was the predominant material before 1968, along with small portions of concrete and asbestos cement pipes. However, from 1968 to 1985, ductile iron was more popular compared to other materials in this network. A mixture of materials has been used from 1985 to date, with PVC having the highest contribution followed by concrete material. The distribution of different materials based on the installation year is provided in the given chart (Figure 0.29).

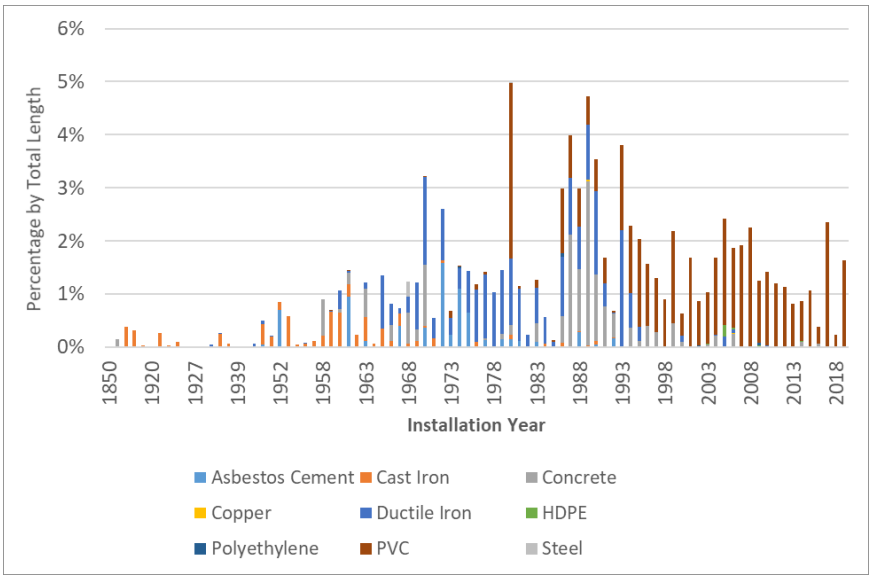


FIGURE 0.29 - PERCENTAGE OF EACH MATERIAL PER TOTAL LENGTH BASED ON INSTALLATION YEAR (REGION OF WATERLOO - INVENTORY)

Regarding diameters, pipes are more normally distributed within this network. Therefore, different diameter ranges can be seen in the given graph, from 100 mm to 1200 mm. For instance, pipes with the size of 450 mm account for almost 30% of the entire network, and ductile iron and PVC are the primary material for this size. For smaller pipes like 150-mm and 200-mm pipes, however, PVC is the predominant material. As the size increases, concrete pipe seems to be more desirable within the Region of Waterloo network (Figure 0.30).

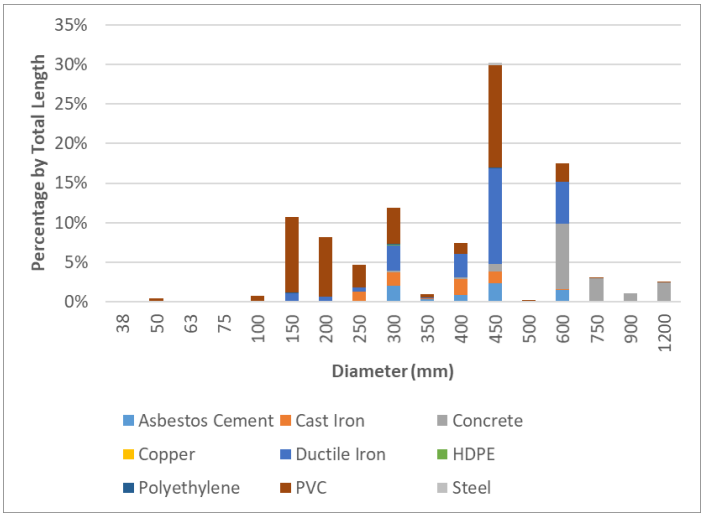


FIGURE 0.30 - PERCENTAGE OF EACH MATERIAL BASED ON SIZE AND TOTAL LENGTH (REGION OF WATERLOO)

Break File

The final excel file for the Region of Waterloo included only 292 failures record. Cast iron accounts for almost 40% of the total failure in the network, followed by ductile iron and PVC pipes with 33% and 20% contribution. The majority of failures are provided without cause and nature of failures. However, from the available information, most failure for cast iron pipes is related to joint and fitting failures in this network. For ductile iron, on the other hand, joint failure and circumferential failure were more frequent. The given figure provides the distribution of different materials based on the total number of failures and the nature of failures (Figure 0.31).

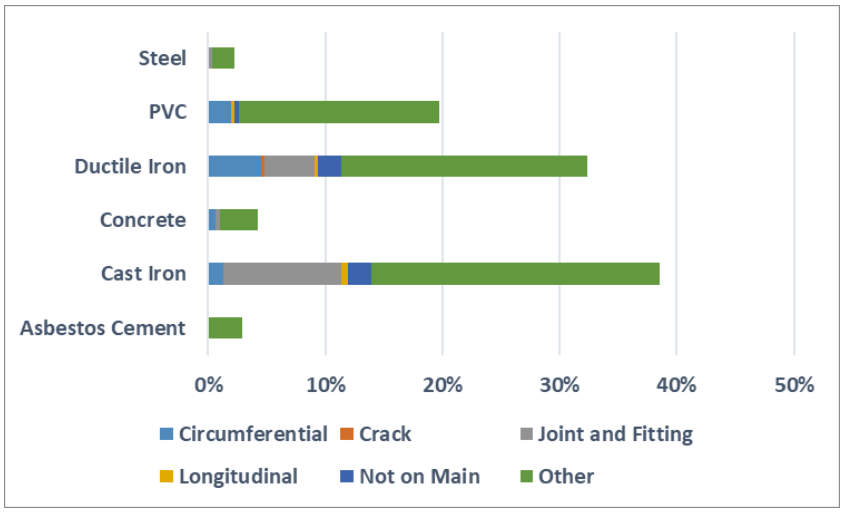


FIGURE 0.31 - PERCENTAGE OF EACH MATERIAL IN THE NETWORK AND THEIR CORRESPONDING FAILURES (BREAK FILE – REGION OF WATERLOO)

Failures for the Region of Waterloo were collected from 1987 to 2019. Figure 0.32 depicts the number of failures in different years based on different materials.

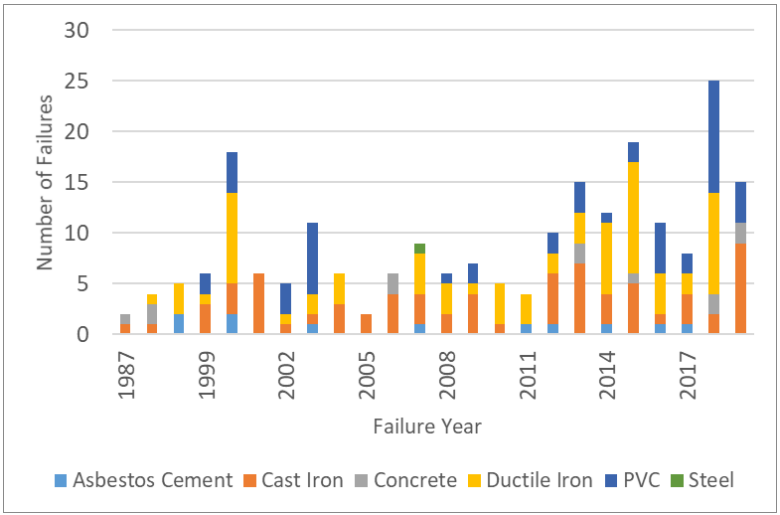


FIGURE 0.32 - NUMBER OF FAILURES FOR DIFFERENT MATERIALS IN DIFFERENT YEARS (REGION OF WATERLOO)

Region of Durham – Ontario

Inventory File

The region of Durham is situated in the Ontario province, Canada. With a population of around 645 thousand and a land area of above 2,523 square kilometers, this network consists of 22,414 pipes. These pipes have different attributes such as diameter, material, installation year, ownership, length, lining status, lining material, lining year, surface type, protection status, protection year, and status. It should be noted that the total length of this network is above 2,638 kilometers. The given table indicates more information regarding these attributes (TABLE 0.7).

TABLE 0.7~~0.7~~ - AVAILABLE ATTRIBUTES WITHIN REGION OF DURHAM INVENTORY DATASET

Attribute	Unit	Range/Values
Diameter	mm	25 - 2100
Material	Text	AC, CI, CON, CO, DI, PE, PVC
Installation Year	Year	1900 - 2020
Ownership	Binary	Yes, No

Length	m	0.15 – 4169.73
Lining Status	Binary	Yes, No
Lining Material	Text	CIP, CM, Unlined
Lining Year	Year	1970 - 2019
Surface Type	Text	Asphalt, Concrete, Creek, Ease, Field, Grass, Gravel, Rail, Stone
Protection Status	Binary	Yes, No
Protection Year	Year	1985 - 2019
Status	Text	Active, Inactive

The predominant material in this network is cast iron, which accounts for almost 50% of the total length. However, ductile iron and PVC are other frequently installed materials in this network. Ductile iron has 33.71% and PVC 11.93% contribution to the total length of pipes (Figure 0.33).

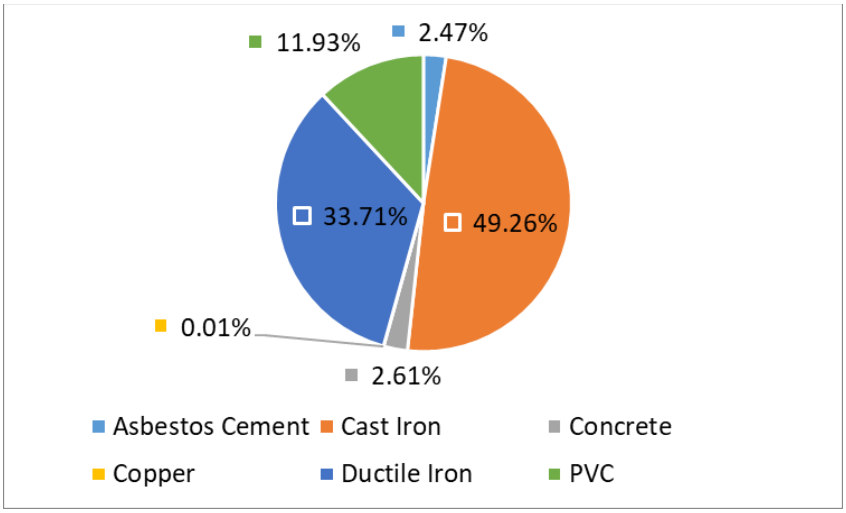


FIGURE 0.33 - DIFFERENT TYPES OF MATERIALS WITHIN REGION OF DURHAM INVENTORY BASED ON THE TOTAL LENGTH

The given figure depicts the distribution of each material installed in different years. As can be seen from 1905 to the 60s, cast iron was the most prevalent type of material in Durham. However, from the 60s to mid-80s, ductile iron was installed more frequently in this network.

Since then, PVC has become the primary material in this network. Interestingly, the number of installations declined significantly from 1982 to date (Figure 0.34).

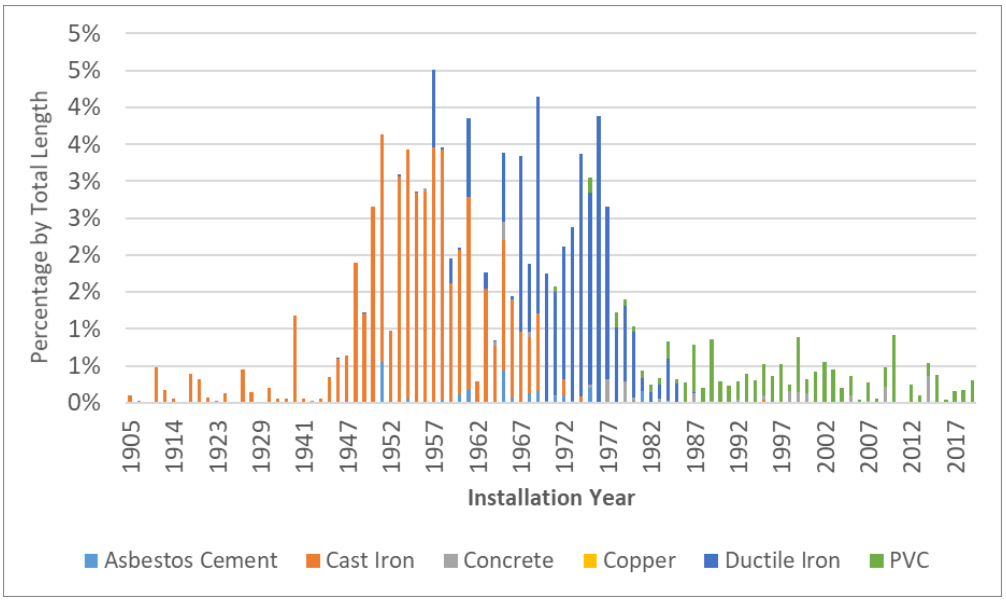


FIGURE 0.34 - PERCENTAGE OF EACH MATERIAL PER TOTAL LENGTH BASED ON INSTALLATION YEAR (REGION OF DURHAM - INVENTORY)

150-mm pipes account for almost 60% of the network, with cast iron being the most frequent. Other sizes such as 200 and 300 pipes only make up over 10% of the total length of pipes in this network. Figure 0.35 shows how different diameters are distributed within the Durham utility.

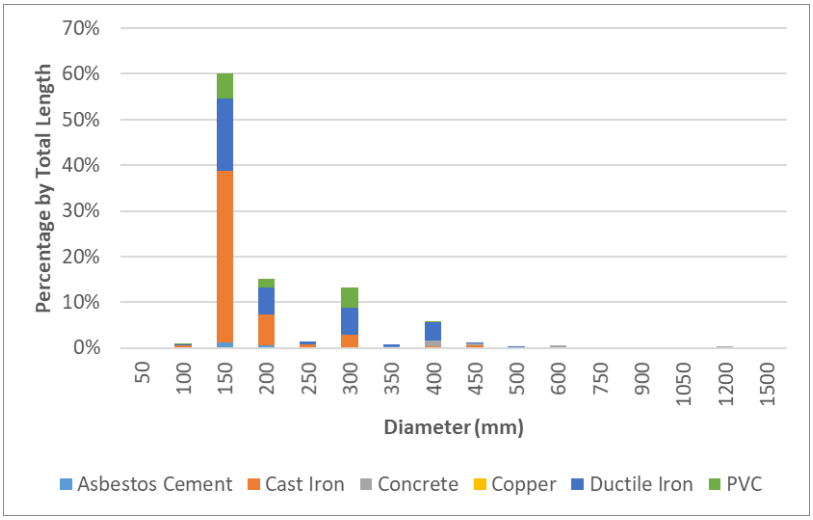


FIGURE 0.35 - PERCENTAGE OF EACH MATERIAL BASED ON SIZE AND TOTAL LENGTH (REGION OF DURHAM)

Break File

The final cleaned dataset includes 6,578 failure records. Among these failures, cast iron is the material the experienced the highest number of failures in the Region of Durham. This material solely accounts for more than 50% of the recorded failures. Circumferential failure is the predominant type of failure for cast iron pipes, followed by the hole and joint-related failures. Ductile iron also makes up 30% of failures in this network, and hole is the primary type of failure for this material. The given bar chart shows the contribution of each material to total failures based on the number of failures and type of failures (Figure 0.36).

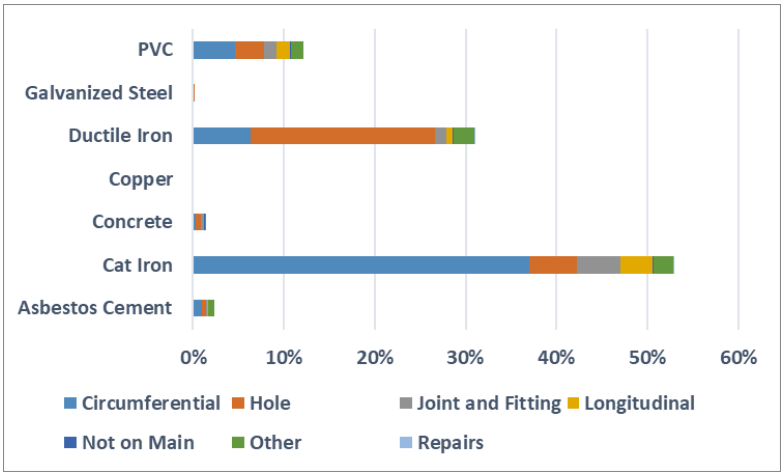


FIGURE 0.36 - PERCENTAGE OF EACH MATERIAL IN THE NETWORK AND THEIR CORRESPONDING FAILURES (BREAK FILE – REGION OF DURHAM)

As shown in the given histogram chart, the number of failures in this network experienced a peak in 1994, then leveled off and declined steadily. Meanwhile, the given graph shows that cast iron pipes are experiencing a significant number of failures, as previously discussed (Figure 0.37).

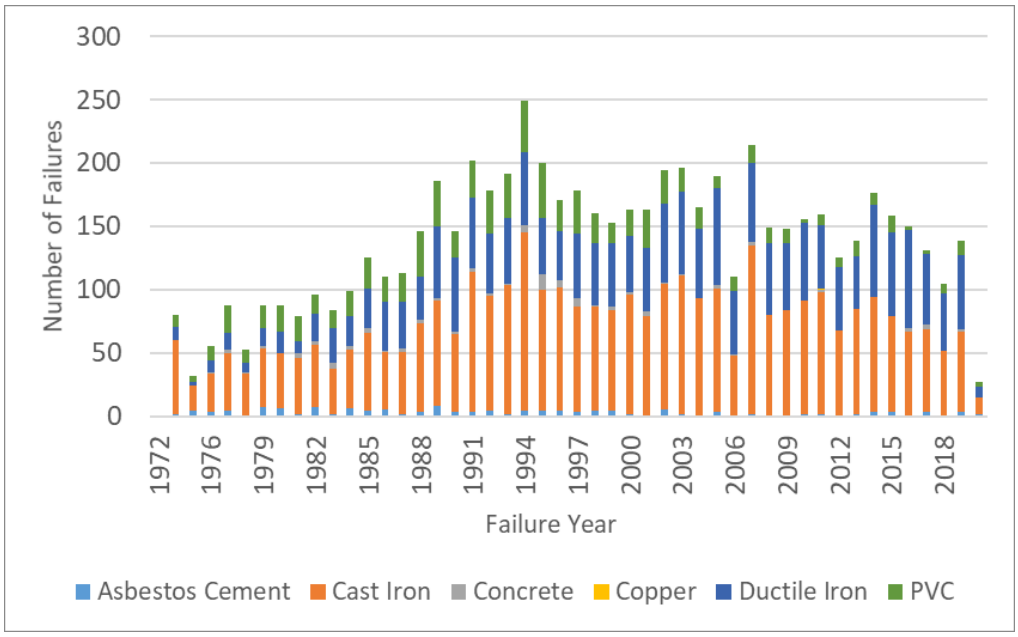


FIGURE 0.37 - NUMBER OF FAILURES FOR DIFFERENT MATERIALS IN DIFFERENT YEARS (REGION OF DURHAM)

Calgary – Alberta

Inventory File

This utility is among the largest networks in this study. Calgary is a large metropolitan located in the Alberta province, Canada. With over 1,300,000 inhabitants, this city is the 4th largest city in Canada, based on the 2016 census. The final dataset of Calgary includes 55,561 pipes, with different attributes such as diameter, material, installation year, length, dead-end, average soil resistivity, break number, and break rate. The total length of this network is approximately 5,277 kilometers. The given table lists all attributes and the range of their values (TABLE 0.8).

TABLE 0.8~~0.8~~ - AVAILABLE ATTRIBUTES WITHIN CALGARY INVENTORY DATASET

Attribute	Unit	Range/Values
Diameter	mm	12 - 3000
Material	Text	AC, CI, CON, CO, DI, PE, PVC, PVCF, ST

Installation Year	Year	1900 - 2019
Length	m	0.05 – 993.85
Dead End	Binary	Yes, No
Average Soil Resistivity	ohm-meter	597 - 2500
Break Number	Number	0 - 28
Break Rate	Rate/m	0 – 99.98

PVC pipe with 54.36% of the total length of this network has the highest contribution. Ductile iron and cast iron are the following materials in this network, which account for 20.12% and 15.06%, respectively. Around 10% of the network belongs to concrete, steel, Polyethylene, and asbestos cement pipes (Figure 0.38).

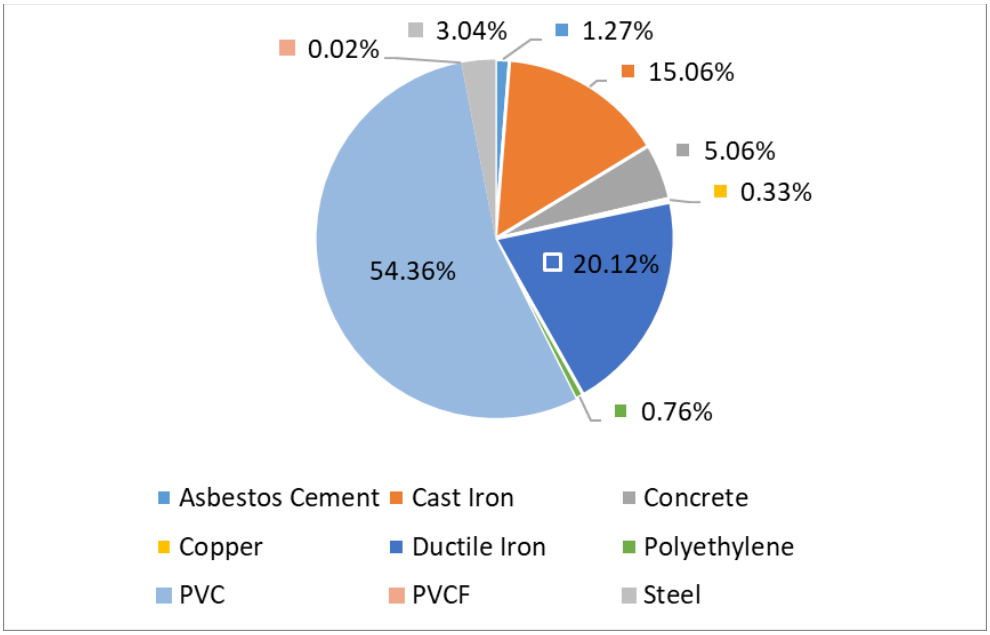


FIGURE 0.38 - DIFFERENT TYPES OF MATERIALS WITHIN CALGARY INVENTORY BASED ON THE TOTAL LENGTH

It is clear that at the beginning of the 20th century, a significant number of cast iron pipes were installed. Then the number of installations declined until 1945, followed by an abrupt increase. From 1945 to 1965, cast iron was the primary material installed within this network. With the introduction of ductile iron, this material then became the predominant type in Calgary until 1980. Ever since, like other utilities, PVC has gained more popularity among other materials (Figure 0.39).

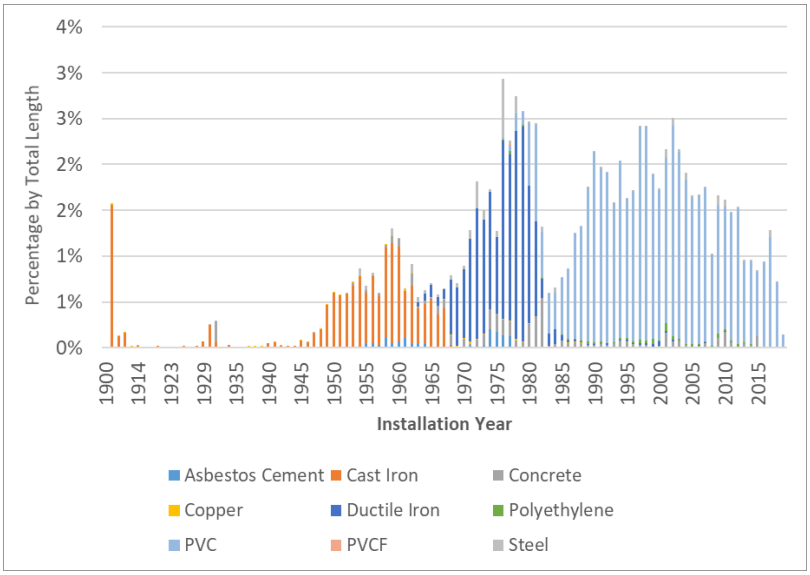


FIGURE 0.39 - PERCENTAGE OF EACH MATERIAL PER TOTAL LENGTH BASED ON INSTALLATION YEAR (CALGARY - INVENTORY)

Turning to diameter, pipes from 150 mm to 300 mm are the majority of the network. Seemingly, with increasing the size of the diameter use of concrete and steel pipes has become more popular. 150-mm pipes with more than 28% of contribution to total length are the mostly employed size.

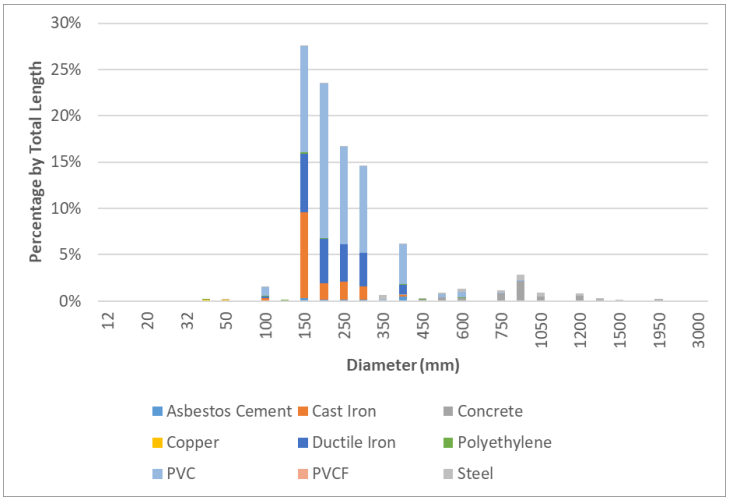


FIGURE 0.40 - PERCENTAGE OF EACH MATERIAL BASED ON SIZE AND TOTAL LENGTH (CALGARY)

Break File

After preparing and cleaning the dataset, broken pipes included 36,396 segments. Again, PVC and cast iron are the materials that underwent the most failures in the Calgary network, with

37% and 34% contribution (Figure 0.41). Circumferential failure is the primary type of failure for cast iron pipes. For PVC and ductile iron, however, the hole is the paramount nature of the failure. It should be mentioned that these materials experienced a significant number of failures related to joint and fittings. Furthermore, this dataset includes break number, failure date, soil type, corrosion degree, soil condition, status, material, diameter, failure type, failure cause, coating status, protection status, and anode type as explanatory variables.

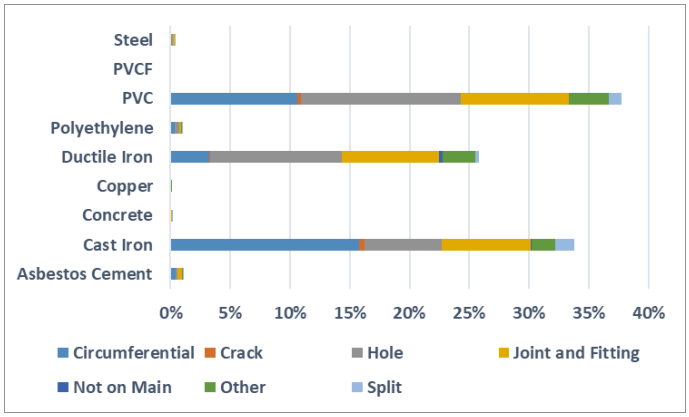


FIGURE 0.41 - PERCENTAGE OF EACH MATERIAL IN THE NETWORK AND THEIR CORRESPONDING FAILURES (BREAK FILE – CALGARY)

Looking at the given graph reveals that the number of failures increased significantly from 1956 to the mid-'80s, with a peak in 1982. Then, the rate of failures declined steadily from 1986 to 2006, then almost leveled off. Figure 0.42 indicates the number of failures in each year based on the materials.

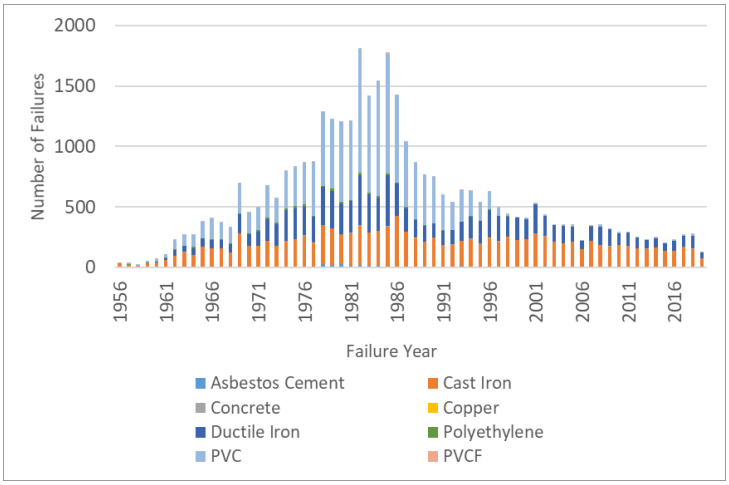


FIGURE 0.42 - NUMBER OF FAILURES FOR DIFFERENT MATERIALS IN DIFFERENT YEARS (CALGARY)

Vancouver – British Columbia

Inventory File

With a population of more than 2,463,000, Vancouver is the third-largest city in Canada, located in British Columbia province. This city has a land area of 2,878 square kilometers, and the inventory file of this network consists of 67,522 pipe segments. The total length of this network, including all pipes, is around 1,626 kilometers. Moreover, different attributes were provided with the excel file of Vancouver, including diameter, material, installation year, length, status, ownership, service type, coating material, and lining material. The list of these attributes and their values is provided in the given table (TABLE 0.9).

TABLE 0.9~~0.9~~ - AVAILABLE ATTRIBUTES WITHIN VANCOUVER INVENTORY DATASET

Attribute	Unit	Range/Values
Diameter	mm	20 - 1950
Material	Text	CI, CON, CO, DI, HDPE, PE, PVC, ST
Installation Year	Year	1892 - 2020
Length	m	0.09 - 2210
Status	Text	Active, Inactive
Ownership	Binary	Yes, No
Service Type	Text	Distribution, Facility, Transmission
Coating Material	Text	Coal Tar, Concrete, Epoxy, Foam, PB, Uncoated
Lining Material	Text	CM, Coal Tar, Epoxy, Polyurea, Unlined

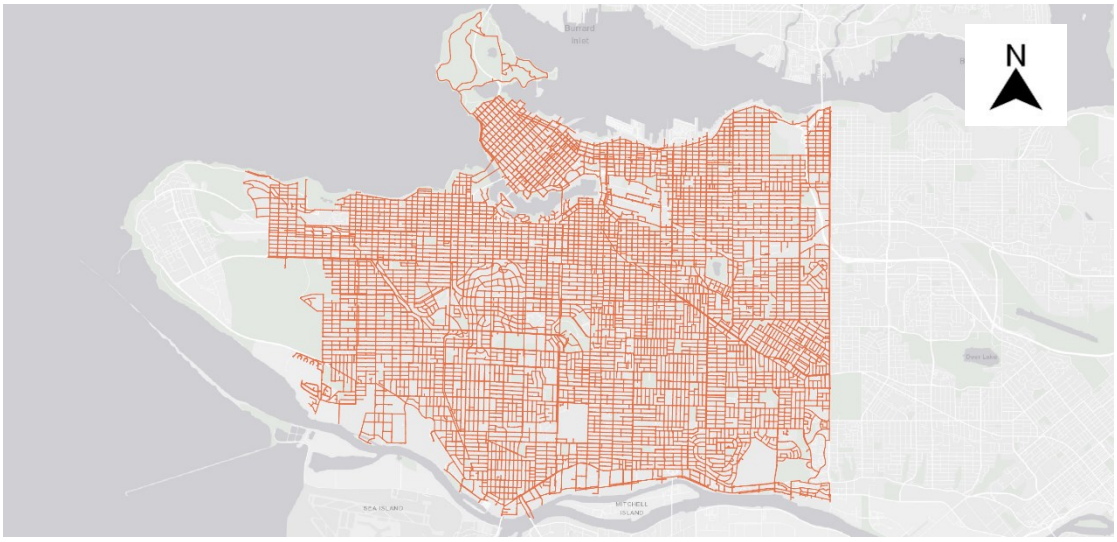


FIGURE 0.43 - VANCOUVER WATER DISTRIBUTION NETWORK (GIS FILE PROVIDED BY CITY OF VANCOUVER)

With 48.54% and 44.12% contribution, cast iron and ductile iron are frequently installed materials within the Vancouver network. Steel pipes account for almost 5.95 of the total length of this network, and other materials make up just small portions of the total length. The given pie chart provides more information about all materials installed in this utility (Figure 0.44).

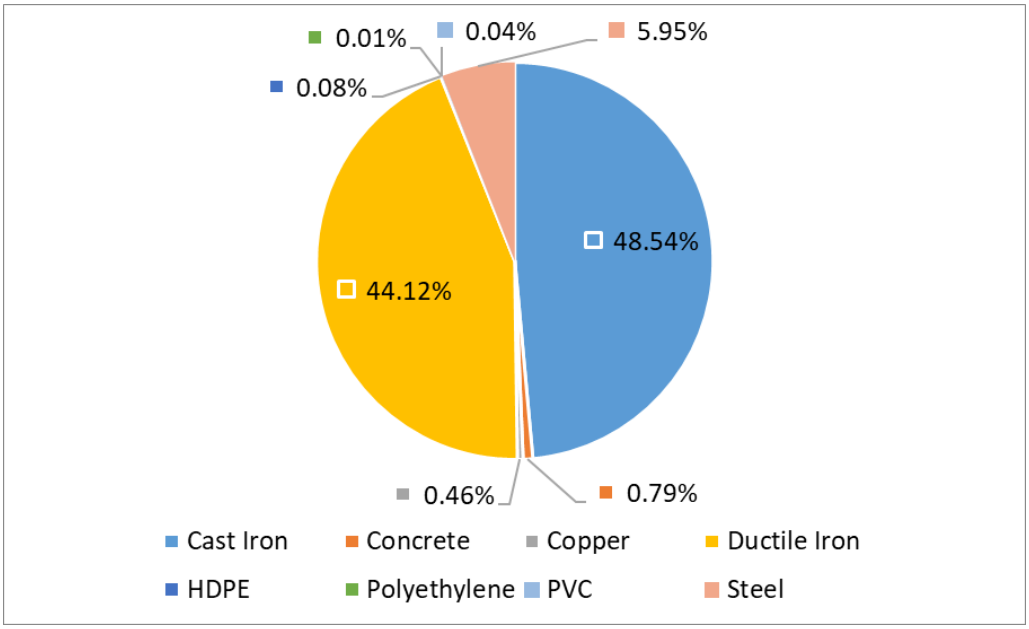


FIGURE 0.44 - DIFFERENT TYPES OF MATERIALS WITHIN VANCOUVER INVENTORY BASED ON THE TOTAL LENGTH

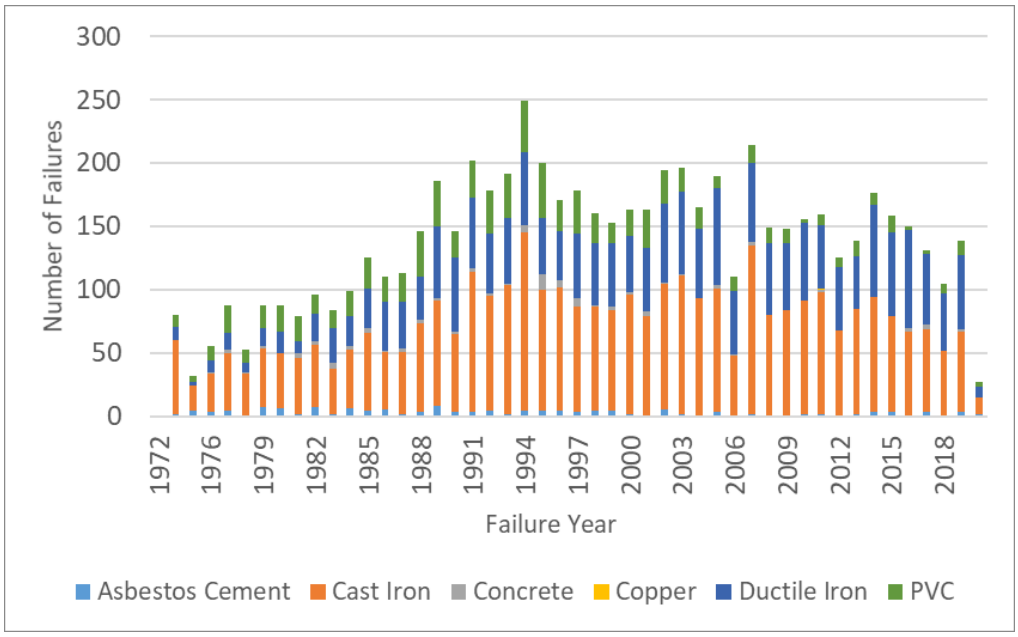


FIGURE 0.45 - PERCENTAGE OF EACH MATERIAL PER TOTAL LENGTH BASED ON INSTALLATION YEAR (VANCOUVER - INVENTORY)

Smaller pipes are more prevalent in this network than larger diameters. For instance, 150-mm and 200-mm pipes with almost 70% of the total length have the highest contribution. For the former, cast iron seems to be more popular in this network. However, for the latter, ductile iron was used more frequently. Furthermore, 300-mm pipes account for almost 20% of the total length in this utility. The given charts indicate the distribution of each material within this network based on the installation year and size of the pipes (Figure 0.45; Figure 0.46).

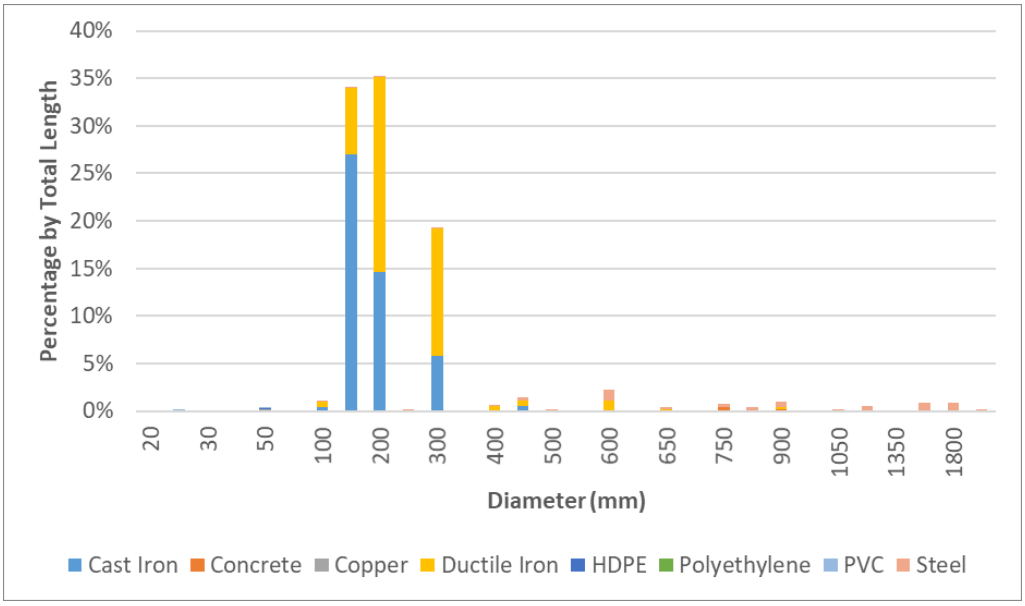


FIGURE 0.46 - PERCENTAGE OF EACH MATERIAL BASED ON SIZE AND TOTAL LENGTH (VANCOUVER)

Break File

The final cleaned file of recorded failure only included 927 pipes, most of which are cast iron, about 90%. The majority of failures for cast iron pipes are circumferential failures and a small portion related to joint and fitting failures. Failure date, service type, diameter, ownership, length, status, soil type, joint type, pipe depth, material, installation year, failure cause, and failure type are the attributes provided by the city of Vancouver. The given chart shows the percentage of each material based on the number of failures recorded in this network (Figure 0.47). Another figure also depicts the distribution of each material based on the number of failures in different years (Figure 0.48). Failure was provided from 2009 to 2020 for this network.

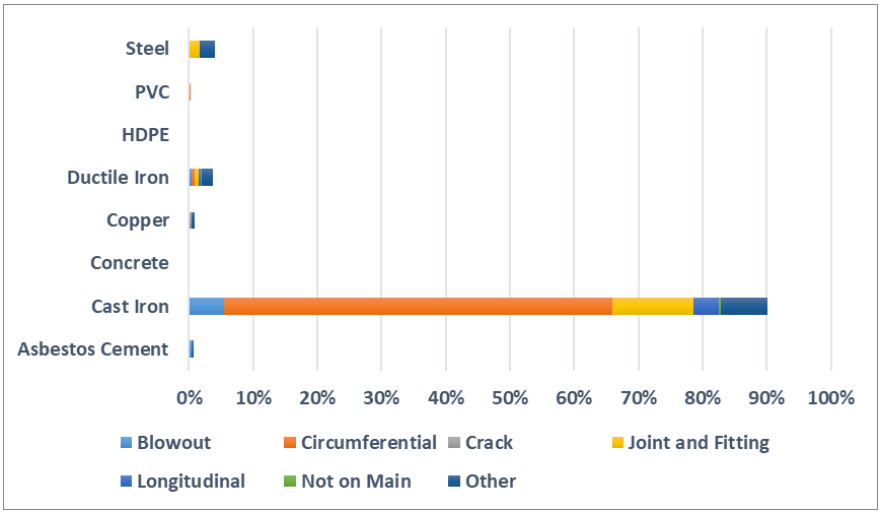


FIGURE 0.47 - PERCENTAGE OF EACH MATERIAL IN THE NETWORK AND THEIR CORRESPONDING FAILURES (BREAK FILE – VANCOUVER)

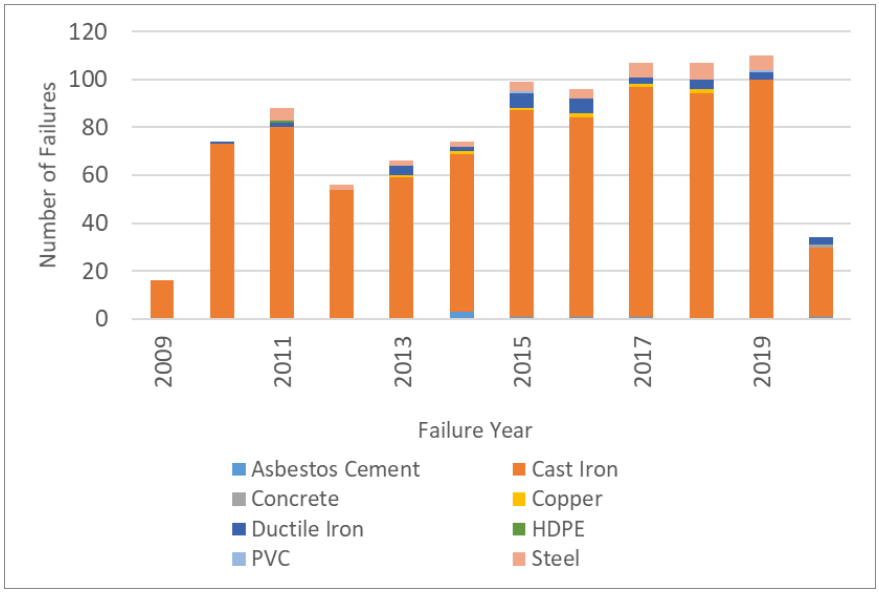


FIGURE 0.48 - NUMBER OF FAILURES FOR DIFFERENT MATERIALS IN DIFFERENT YEARS (VANCOUVER)

Victoria – British Columbia

Inventory File

Victoria is another network located in British Columbia province with a population of about 367 thousand with a land area of approximately 696 square kilometers. The inventory file of this network after cleaning includes 3,319 pipes with different input variables. These attributes and their ranges are provided in the given table (TABLE 0.10). Moreover, according to the available information, the total length of this network is approximately 332 kilometers.

TABLE 0.10~~0.10~~ - AVAILABLE ATTRIBUTES WITHIN VICTORIA INVENTORY DATASET

Attribute	Unit	Range/Values
Diameter	mm	19 - 990
Material	Text	AC, CI, CO, DI, GI, HDPE, PE, PVC, PVCO, ST
Installation Year	Year	1888 - 2016
Length	m	0.15 - 716
Status	Text	Active, Inactive
Ownership	Binary	Yes, No
HGL	Text	72 - 116

Roughness	μ	26 - 140
Lining Material	Text	CM, EPOXY, HDPE, Unlined
Lining Status	Binary	Yes, No

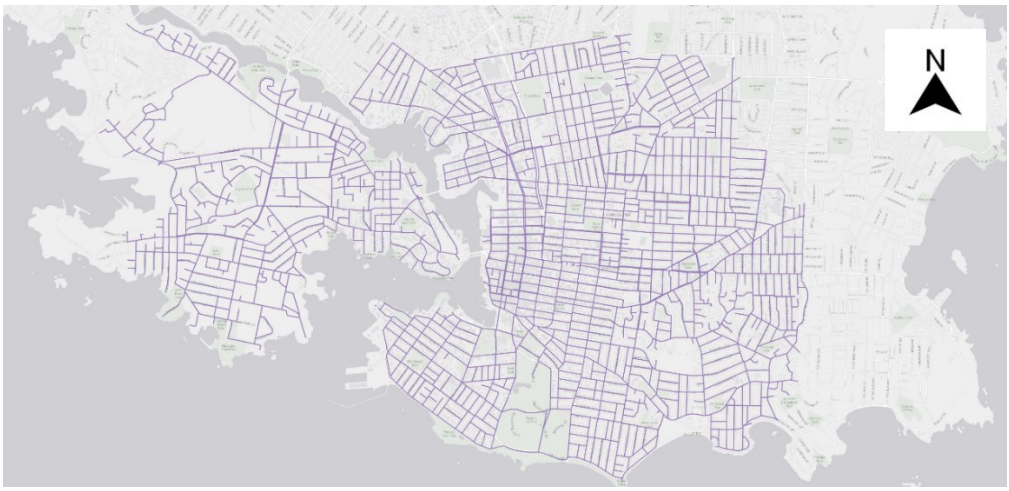


FIGURE 0.49 - VICTORIA WATER DISTRIBUTION NETWORK (GIS FILE PROVIDED BY CITY OF VICTORIA)

Cast iron and ductile iron account for a significant proportion of total pipes installed in this network, with 45.94% and 40.22% contribution to the total length. PVC with a 7.83% contribution is the following material. The given pie chart shows all materials within the inventory file based on the percentage of total length (Figure 0.50).

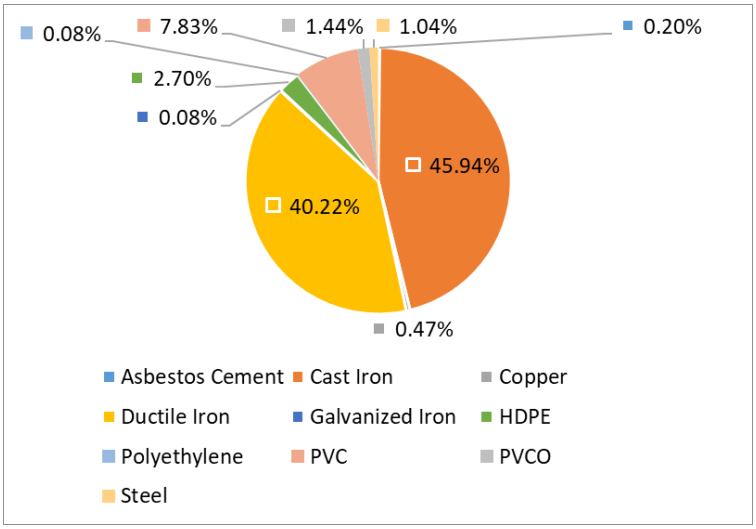


FIGURE 0.50 - DIFFERENT TYPES OF MATERIALS WITHIN VICTORIA INVENTORY BASED ON THE TOTAL LENGTH

Figure 0.51 represents the distribution of different materials in terms of installation in different years. From 1888 to 1961, cast iron was the predominant type of material, same as other networks. However, since then, ductile iron has become the most popular type of pipe in the

Victoria network. Furthermore, according to the available information, 150-mm pipes are the most frequently used in this network. It seems that for larger pipes, HDPE material was employed. Figure 0.52 shows the distribution of different sizes in the Victoria network.

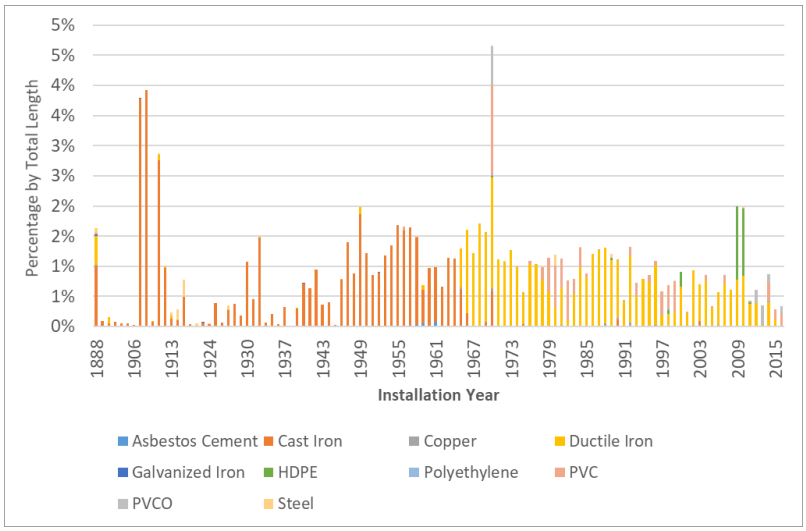


FIGURE 0.51 - PERCENTAGE OF EACH MATERIAL PER TOTAL LENGTH BASED ON INSTALLATION YEAR (VICTORIA - INVENTORY)

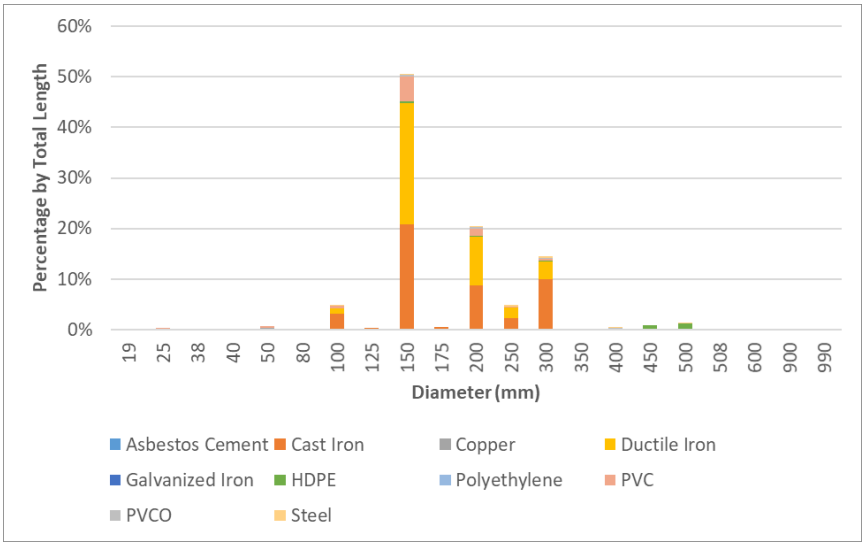


FIGURE 0.52 - PERCENTAGE OF EACH MATERIAL BASED ON SIZE AND TOTAL LENGTH (VICTORIA)

Break File

The failure record of Victoria consists of 977 pipes with different attributes such as failure date, material, diameter, failure type, and failure cause. Almost 57% of these failures are related to

cast iron pipes, with circumferential failure as the most frequent nature of the failure. Other failures reported in this network are split and longitudinal crack. Finally, ductile iron and PVC are the following materials with the highest failure rate after cast iron pipes (Figure 0.53).

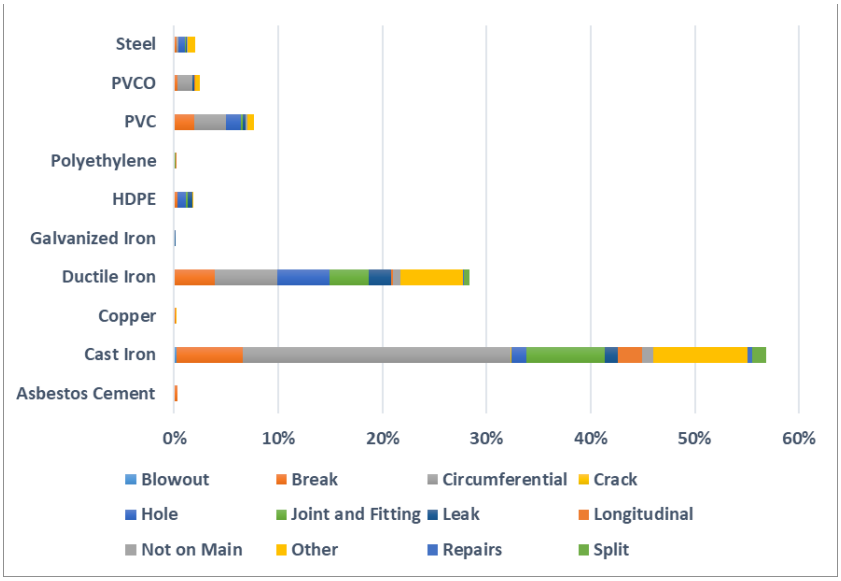


FIGURE 0.53 - PERCENTAGE OF EACH MATERIAL IN THE NETWORK AND THEIR CORRESPONDING FAILURES (BREAK FILE – VICTORIA)

Failure for this network was provided from 1985 to 2019, as shown in the given graph. This graph also shows the distribution of the number of failures based on different materials and years of failure (Figure 0.54).

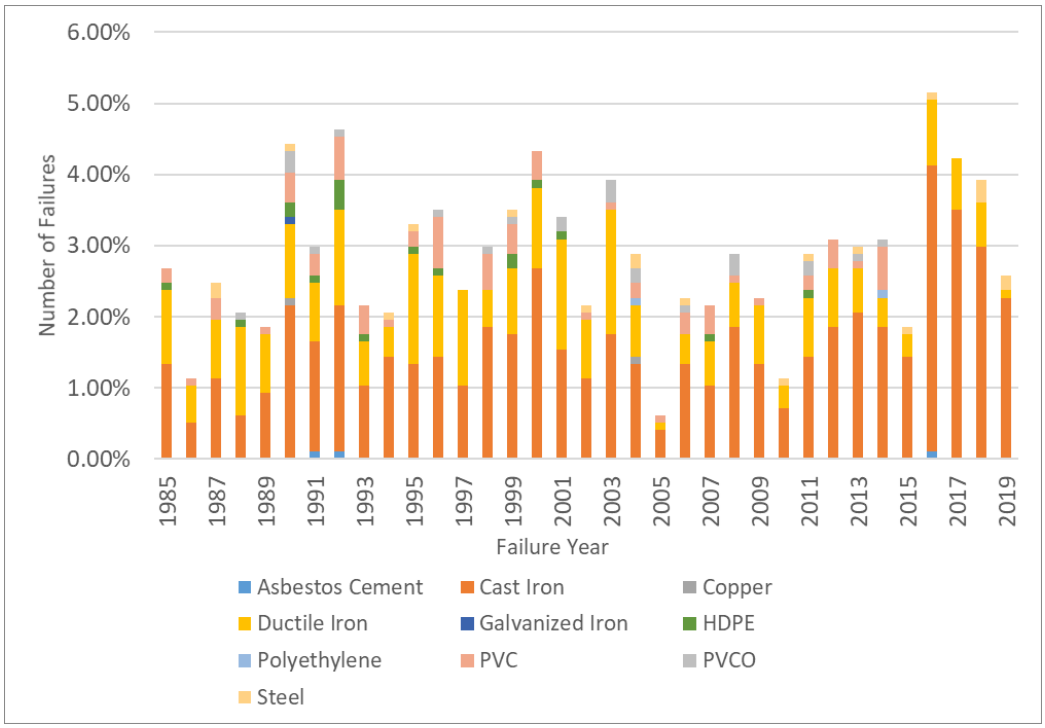


FIGURE 0.54 - NUMBER OF FAILURES FOR DIFFERENT MATERIALS IN DIFFERENT YEARS (VICTORIA)

Halifax– Nova Scotia

Inventory File

Halifax is another utility situated in Nova Scotia province, Canada. Based on the 2016 census, the population of this city is around 403 thousand, with a total land area of 5,490 square kilometers. The final inventory file of this network consists of 14,436 pipes with a total length of approximately 1,566 kilometers. The excel file of this network included a variety of input variables for the modeling process, such as material, diameter, installation year, length, lining material, and lining status. The range of these attributes is listed in the given table (TABLE 0.11).

TABLE 0.110.11 - AVAILABLE ATTRIBUTES WITHIN HALIFAX INVENTORY DATASET

Attribute	Unit	Range/Values
Diameter	mm	19 - 1500
Material	Text	AC, Brass, CI, CON, CO, CLPE, DI, GST, HDPE, PVC, SST, ST
Installation Year	Year	1856 - 2019

Length	m	0.03 - 3620
Lining Material	Text	CM, Polyurea, Unlined
Lining Status	Binary	Yes, No

Ductile iron seems to have been the most frequently installed material within this network, which accounts for almost 60% of the total length. Same as other utilities, cast iron has also been a popular pipe type with a 27.63% contribution. PVC and concrete are other types of materials used in this network, with over 5% contribution for each (Figure 0.55).

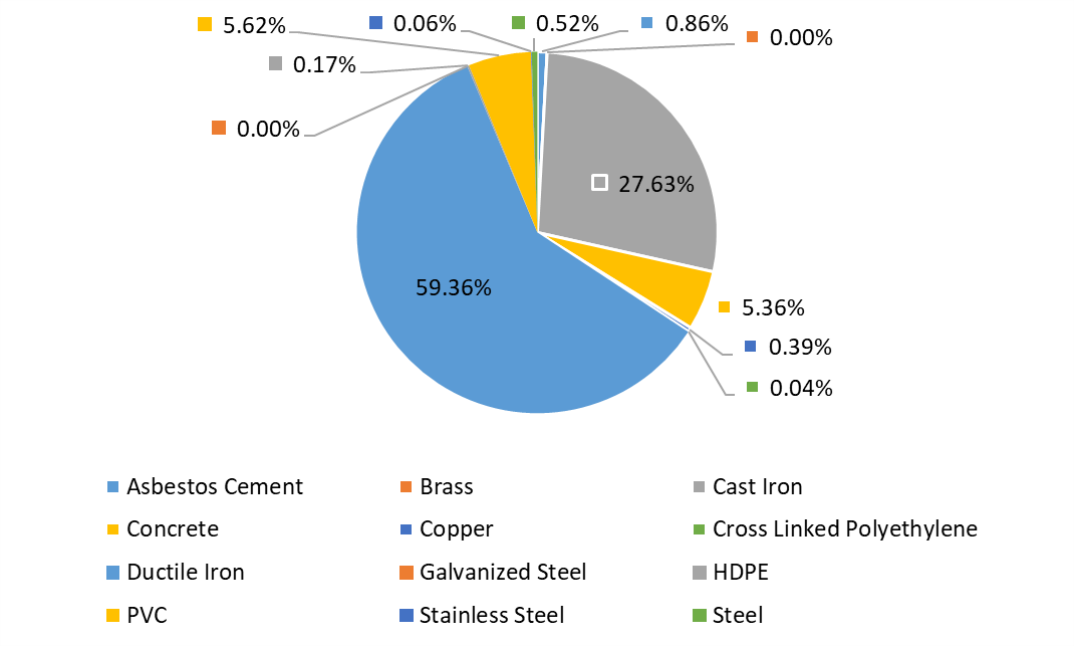


FIGURE 0.55 - DIFFERENT TYPES OF MATERIALS WITHIN HALIFAX INVENTORY BASED ON THE TOTAL LENGTH

Figure 0.56 shows the distribution of installation of each material within Halifax utility. As shown in the chart, cast iron was the predominant material from the beginning until the early '70s. Ever since, however, ductile iron has played a critical role in this network, with the highest contribution to the installation. Concrete and PVC are other materials used in this network, as can be noticed from the given bar chart.

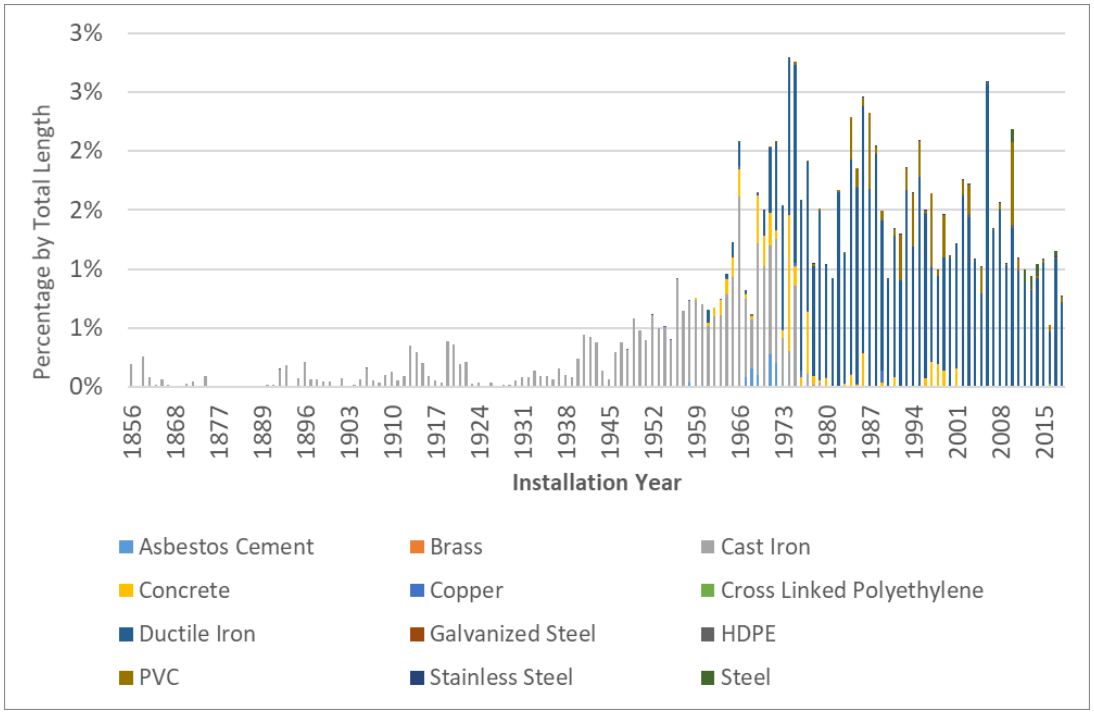


FIGURE 0.56 - PERCENTAGE OF EACH MATERIAL PER TOTAL LENGTH BASED ON INSTALLATION YEAR (HALIFAX - INVENTORY)

Different ranges of diameter have been used in this network. However, smaller pipes have been installed more frequently. For instance, 200-mm pipes were the most popular size in this network, with over 32% of the total length. 150-mm and 300-mm pipes are in the following positions, which account for 20% and 15% of the total length in the Halifax network. The given figures indicated the distribution of each material based on its length and size.

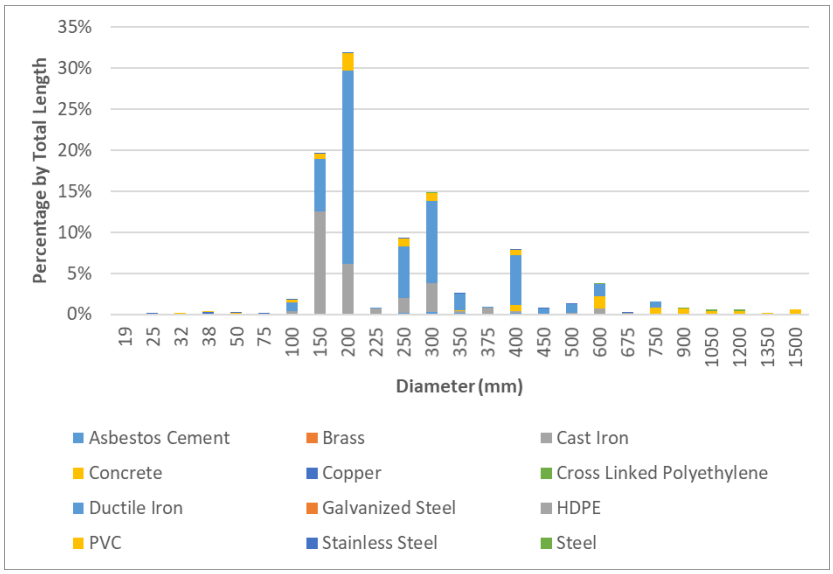


FIGURE 0.57 - PERCENTAGE OF EACH MATERIAL BASED ON SIZE AND TOTAL LENGTH (HALIFAX)

Break File

After cleaning, the excel file related to broken pipes included 6,381 pipes, with cast iron as the pipe that experienced the highest number of failures, with almost 90% of total records. Thus, circumferential failure is the most frequent nature of failures for this pipe. Meanwhile, ductile iron with almost 8% of total failures is the following material in this network (Figure 0.58). It should be mentioned that there are only two primary attributes within this file for Halifax, which are failure date and failure type. Other features such as material and diameter were extracted from the inventory file for this city.

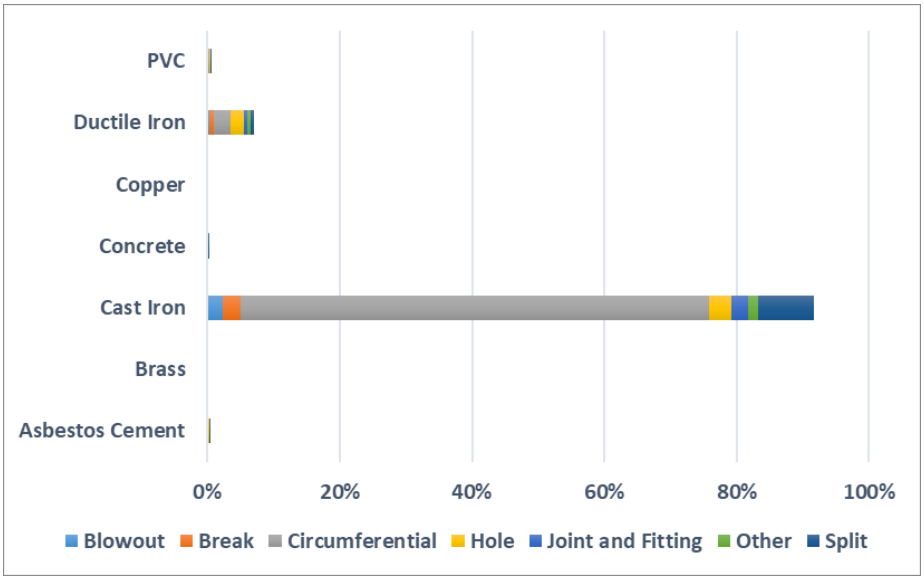


FIGURE 0.58 - PERCENTAGE OF EACH MATERIAL IN THE NETWORK AND THEIR CORRESPONDING FAILURES (BREAK FILE – HALIFAX)

Finally, Figure 0.59 shows the number of failures that occurred in different years. As can be seen, cast iron pipes experienced a higher failure rate than other materials in every individual year.

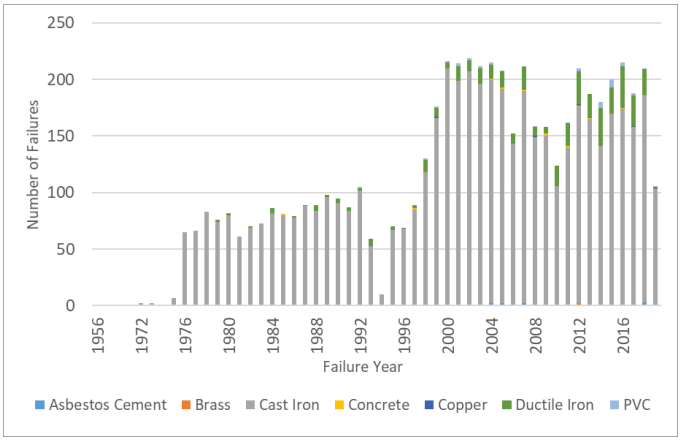


FIGURE 0.59 - NUMBER OF FAILURES FOR DIFFERENT MATERIALS IN DIFFERENT YEARS (HALIFAX)

St. John's - Newfoundland and Labrador

Inventory File

This network is located in Newfoundland and Labrador province, Canada. The population of this city is around 205 thousand, and its total land area is approximately 445 square kilometers. The final inventory file of this network includes 8,983 pipe segments, with a total length of 628 kilometers. Diameter, material, installation year, length, roughness are among the attributes provided by this city.

TABLE 0.120.12 - AVAILABLE ATTRIBUTES WITHIN ST. JOHN'S INVENTORY DATASET

Attribute	Unit	Range/Values
Diameter	mm	12 - 1400
Material	Text	AC, CI, CON, CO, CLPE, DI, HDPE, PE, PVC
Installation Year	Year	1892 - 2017
Length	m	0.016 - 4767
Roughness	μ	10 - 150

FIGURE 0.60 – ST. JOHN'S WATER DISTRIBUTION NETWORK (GIS FILE PROVIDED BY CITY OF ST. JOHN'S)

Ductile iron has the highest contribution to this network based on the total length, with almost 46.47% of the entire length. On the other hand, cast iron and PVC account for almost 41.89% and 10.53% of this network, respectively. There are also other materials, distribution of which can be found in the given pipe chart (Figure 0.61).

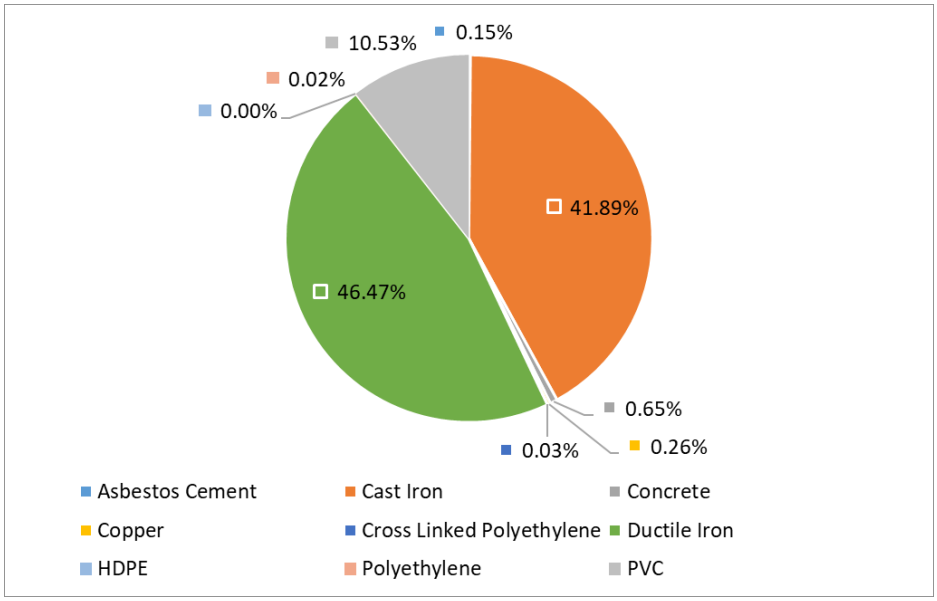


FIGURE 0.61 - DIFFERENT TYPES OF MATERIALS WITHIN ST. JOHN'S INVENTORY BASED ON THE TOTAL LENGTH

The given chart shows the percentage of total length installed in different years since the data collection has started in this network. As shown, from 1892 to 1970, cast iron was predominantly installed in this network. Then, however, ductile iron was the material with more popularity in St.John's, from 1970 to 2008. Since then, PVC has played a vital role in this utility (Figure 0.62).

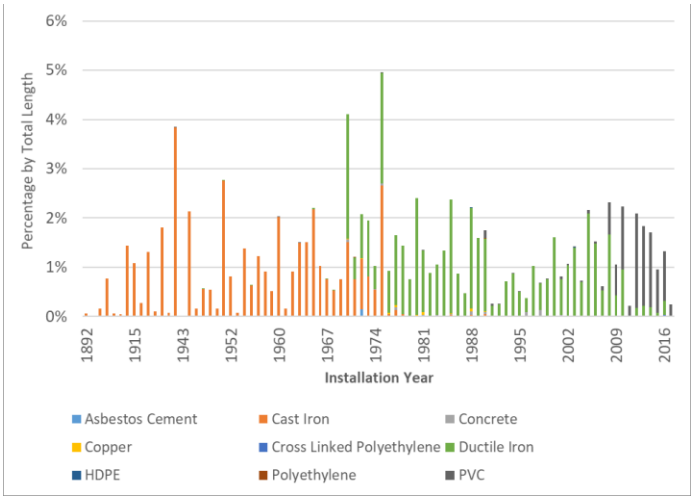


FIGURE 0.62 - PERCENTAGE OF EACH MATERIAL PER TOTAL LENGTH BASED ON INSTALLATION YEAR (ST. JOHN'S - INVENTORY)

Size of 150,200 and 300 mm have been used more frequently in this network compared to other diameters. For 150-mm pipes, cast iron seems to have been more popular than other materials. However, for 200-mm and 300-mm pipes, ductile iron was installed more than other materials. The given bar chart shows the frequency of each material in this network, based on the total length and size of the pipes (Figure 0.63).

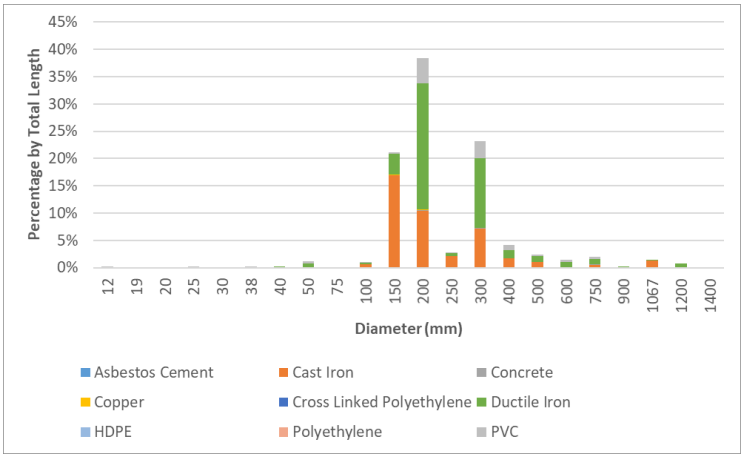


FIGURE 0.63 - PERCENTAGE OF EACH MATERIAL BASED ON SIZE AND TOTAL LENGTH (ST. JOHN'S)

Break File

The failure record of St. John's consists of 1,626 pipes with different attributes such as failure date, material, diameter, failure type, failure cause, and pipe depth. Almost 85% of these failures are related to cast iron pipes, with circumferential and longitudinal failures as the most frequent nature of the failures. Other failures reported in this network are split and longitudinal crack. Finally, ductile iron is the material with the highest failure rate after cast iron pipes, accounting for 10% of the recorded failures (Figure 0.64).

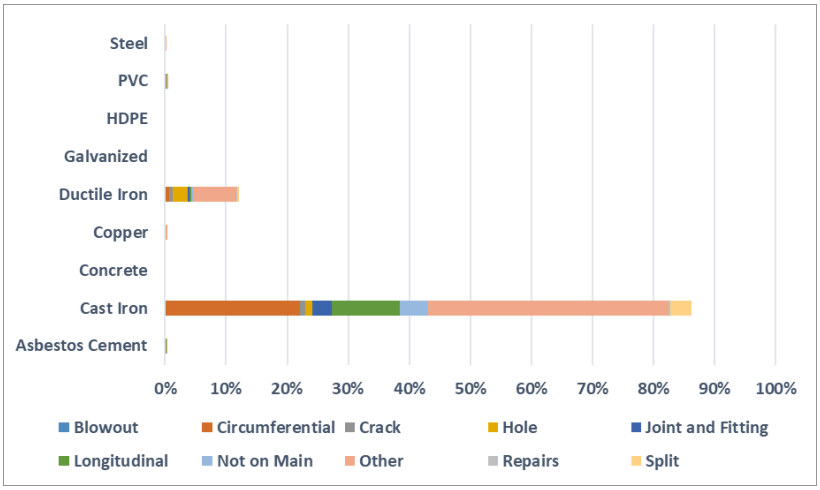


FIGURE 0.64 - PERCENTAGE OF EACH MATERIAL IN THE NETWORK AND THEIR CORRESPONDING FAILURES (BREAK FILE – ST. JOHN'S)

Last but not least is the number of failures that occurred in different years. Failures were reported from 1988 to 2018 for this network. The contribution of each material in this period can be found in the given chart. As previously discussed, cast iron experienced more failures than other materials in this network (Figure 0.65).

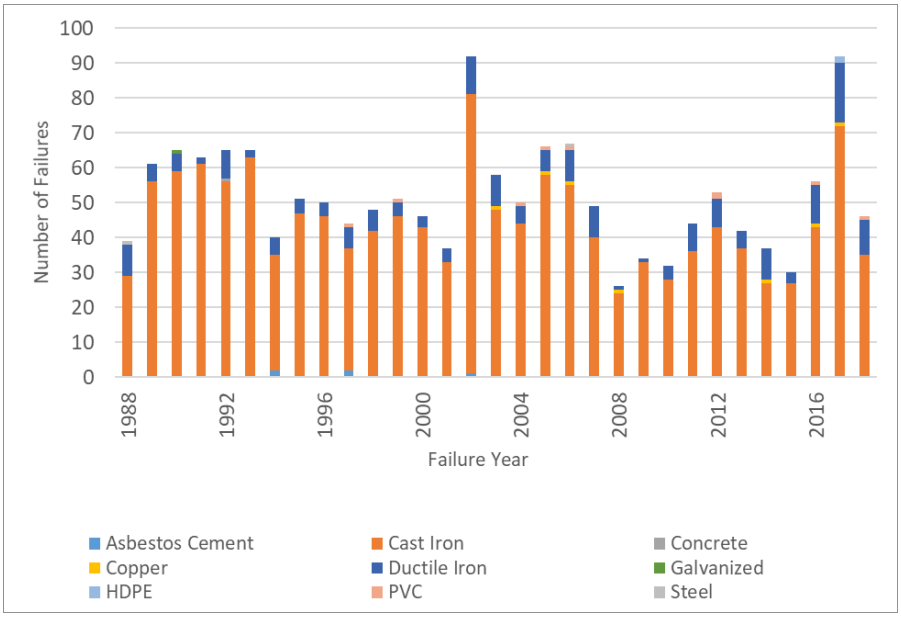


FIGURE 0.65 - NUMBER OF FAILURES FOR DIFFERENT MATERIALS IN DIFFERENT YEARS (ST. JOHN'S)

Barrie - Ontario

Inventory File

Barrie is another city located in Ontario province, Canada. This city has a population of around 212 thousand and a land area of 898 square kilometers. The final inventory file of this utility includes 6,522 pipes with a total length of approximately 749 kilometers. This city provided a range of attributes for conducting this research, including service type, material, diameter, protection status, length, status, casing material, restrained and install year. The range of these attributes can be seen in the given table (TABLE 0.13).

TABLE 0.130.13 - AVAILABLE ATTRIBUTES WITHIN BARRIE INVENTORY DATASET

Attribute	Unit	Range/Values
Service Type	Text	Distribution, Service, Transmission
Material	Text	AC, CI, CON, CO, CLPE, DI, GST, HDPE, PVC, ST
Diameter	mm	19 - 1250

Protection Status	Binary	Yes, No
Length	m	0.1 - 3008
Status	Text	Active, Inactive
Casing Material	Text	Concrete, No casing, Polyethylene, Polystyrene, Steel, StyroFoam, Tunnel
Restrained	Binary	Yes, No
Install Year	Year	1891 - 2019

PVC pipes have been installed more frequently than other materials in this network, which accounts for 54.30% of the total length. Ductile iron with 26.58% contribution is the other material that has been used widely. Cast iron with 11.50% is the third material in this network. It should be noted that there are also other materials, and a percentage of them are listed in the given pie chart (Figure 0.66).

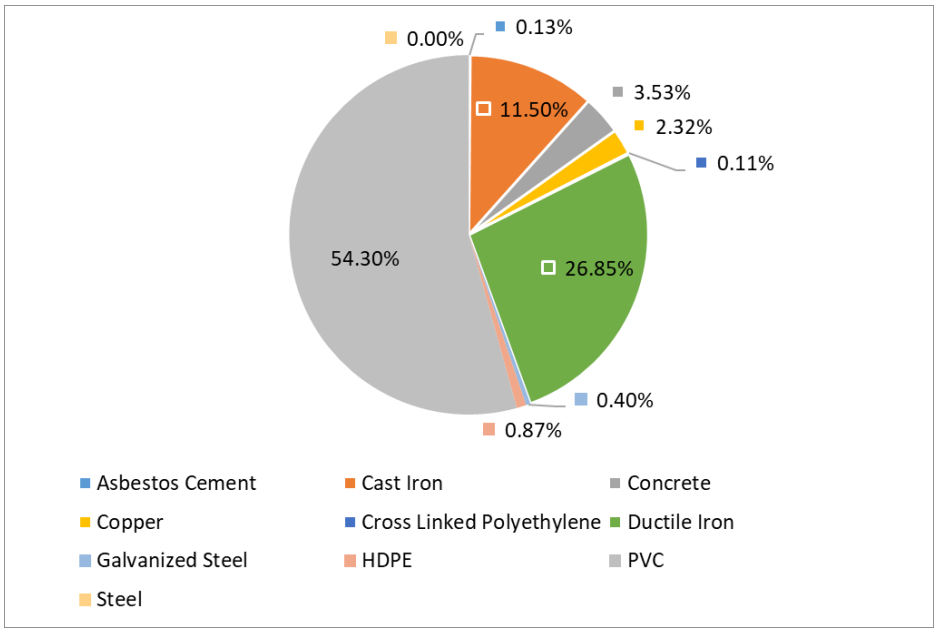


FIGURE 0.66 - DIFFERENT TYPES OF MATERIALS WITHIN BARRIE INVENTORY BASED ON THE TOTAL LENGTH

Figure 0.67 shows the distribution of each material based on the installation year in this network. Like other utilities, cast iron was first the predominant material in Barrie. However, from 1972 to date, ductile iron and PVC have been the most frequently installed materials in this utility.

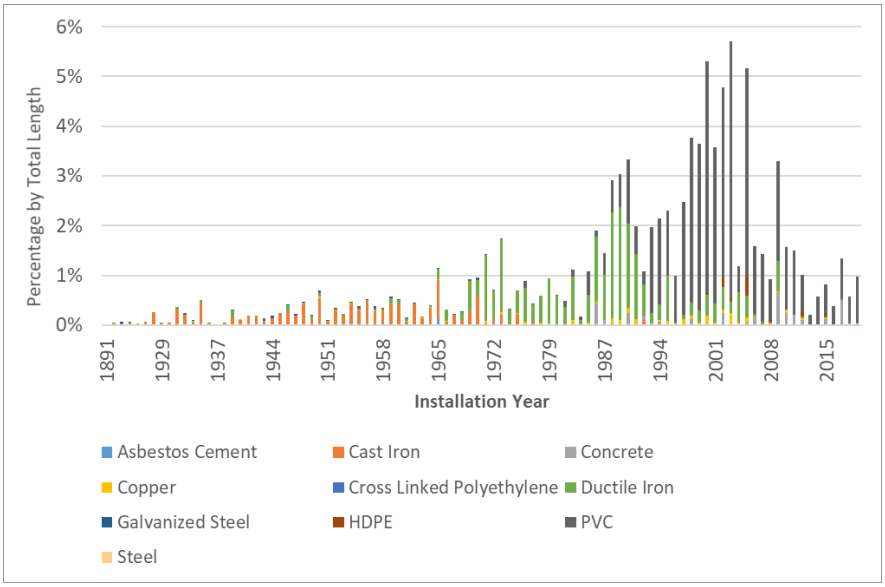


FIGURE 0.67 - PERCENTAGE OF EACH MATERIAL PER TOTAL LENGTH BASED ON INSTALLATION YEAR (BARRIE - INVENTORY)

Similar to other utilities, smaller pipes played a vital role in this network, with 150-mm pipes as the most popular, accounting for almost 45% of total length. Moreover, 200-mm and 300-mm pipes are the following sizes that have been used in this network more frequently than other diameters.

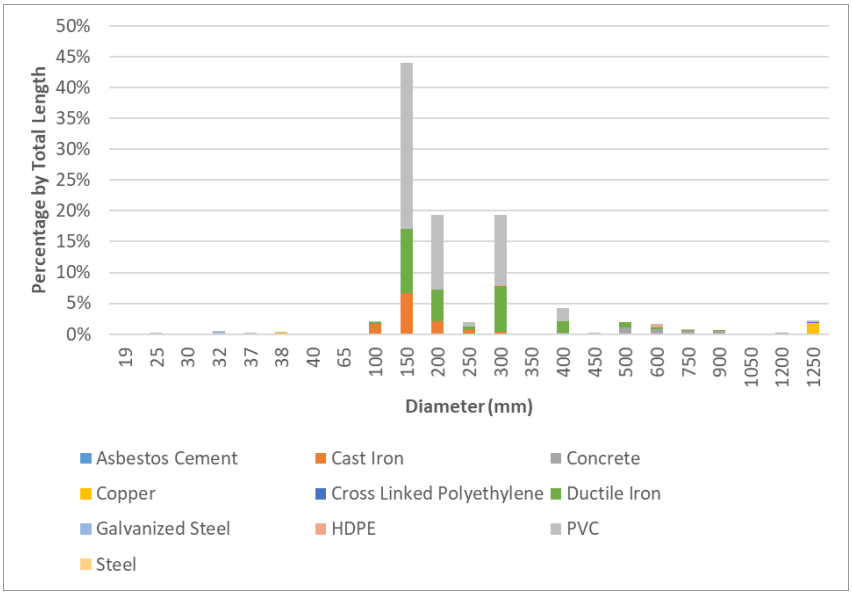


FIGURE 0.68 - PERCENTAGE OF EACH MATERIAL BASED ON SIZE AND TOTAL LENGTH (BARRIE)

Break File

The failure record of this network consists of 1,297 pipe segments with different attributes such as failure date, material, diameter, failure type, failure cause, anode status, break number, and pipe depth. Almost 87% of these failures are related to cast iron pipes, with circumferential and longitudinal failures as the most frequent natures of the failures. Other failures reported in this network are split and hole. Finally, ductile iron is the material with the highest failure rate after cast iron pipes, accounting for 12% of the recorded failures. The given bar chart shows the percentage of each material that failed in this network based on the failure type (Figure 0.69).

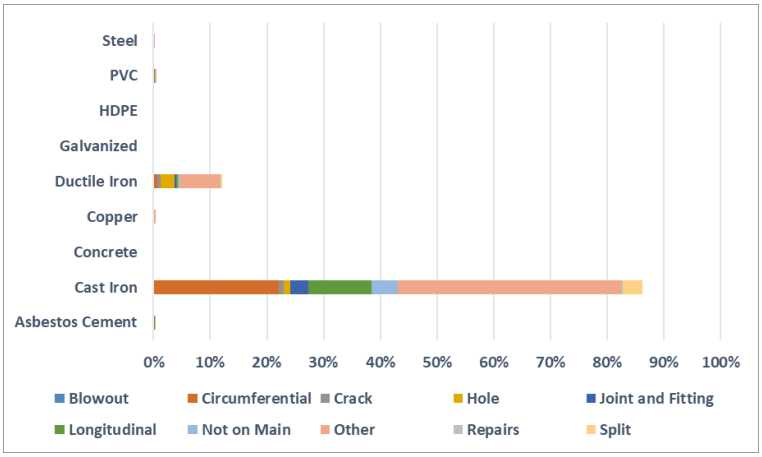


FIGURE 0.69 - PERCENTAGE OF EACH MATERIAL IN THE NETWORK AND THEIR CORRESPONDING FAILURES (BREAK FILE – BARRIE)

Based on the available historical information - 1951 to 2014 - the number of failures increased steadily in this network, with a peak in 2014. Then the number of failures experienced an abrupt decline from 2014 to date. This could be related to applying different practices to maintain the network or directly related to other factors such as age.

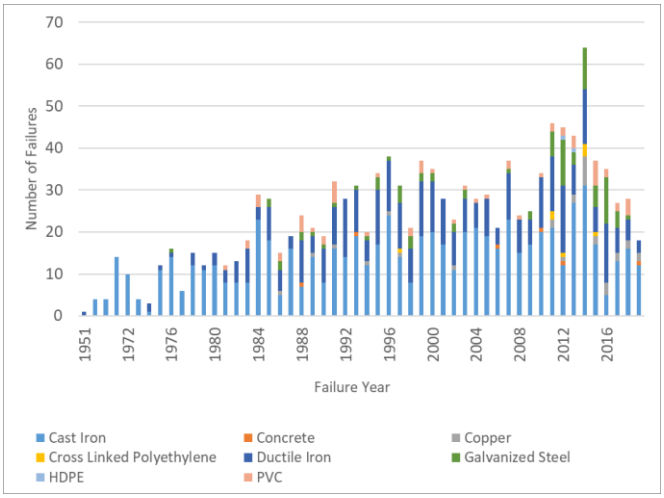


FIGURE 0.70 - NUMBER OF FAILURES FOR DIFFERENT MATERIALS IN DIFFERENT YEARS (BARRIE)

APPENDIX B – ALGORITHMS AND HYPERPARAMETERS

Random forest hyperparameter (Classification and regression)

Random Forest has different parameters that could be tuned based on the desired outcomes. The given figure indicates the entire parameters and their default values based on the Scikit-learn library for random forest classifier (Figure 0.1).

FIGURE 0.1 – RANDOM FOREST CLASSIFIER HYPERPARAMETERS (HTTPS://SCIKIT-LEARN.ORG/STABLE/MODULES/GENERATED/SKLEARN.ENSEMBLE.RANDOMFORESTCLASSIFIER.HTML)

Some of these parameters have remained unchanged during the modeling process. However, a few important ones that may avoid overfitting have been employed and explained in this section.

- ***n_estimators (default = 100):***

This parameter shows the number of trees utilized to learn the base model. As previously mentioned, the output of RF is average or the majority vote of numbers of trees. Thus, finding an appropriate number of trees can be considered the most important step.

- ***Criterion (default = "gini"):***

The criterion function calculates the splitting quality during the learning process. The default is gini for the Gini impurity and entropy for the information gain. Therefore, either of these two methods should be evaluated in order to find the most efficient one.

- ***max_depth (default = None):*** This parameter defines the depth or level of splitting for each tree in the Forest. If it remains as None, the trees are expanded until all nodes reach the maximum impurity or until each node contains fewer samples than defined "min_samples_split."

- ***min_samples_split (default = 2):*** This parameter is typically used for controlling overfitting. The digit number shows how many samples should remain in each node in order to continue the splitting process. If there are fewer samples in each node than the defined value, the splitting process will stop. In this study, this parameter is used for the datasets that contain a higher number of samples

- ***max_features (default = "auto"):.*** This parameter defines the number of variables that are considered during the splitting process. Different values may be allocated based on the availability of different attributes such as auto, sqrt, log2, int, float, and None.

auto = sqrt (n_features)

sqrt = sqrt (n_features)

log2 = log2 (n_features)

None = n_feature

For instance, if max_features is defined as "sqrt" and nine variables are available, the number of features utilized in the analysis would be 3.

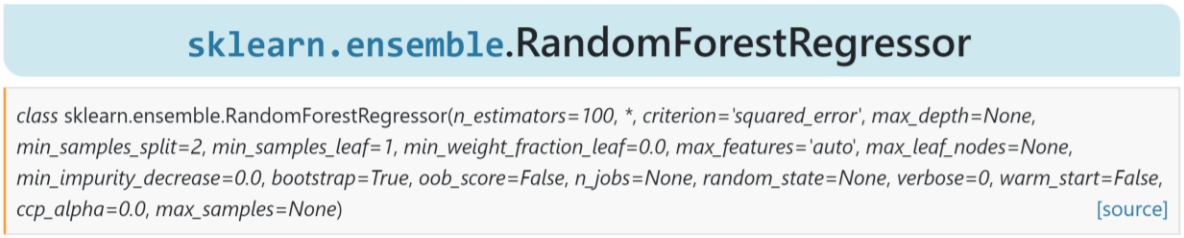


FIGURE 0.2 - RANDOM FOREST REGRESSOR HYPERPARAMETERS (HTTPS://SCIKIT-LEARN.ORG/STABLE/MODULES/GENERATED/SKLEARN.ENSEMBLE.RANDOMFORESTREGRESSOR.HTML)

Figure 0.2 represents different parameters for random forest regressor. Except for criterion that is different for regression problems, other hyperparameters are similar to the random forest classifier.

- ***Criterion (default = "squared_error"):*** This parameter also measures the quality of each split. Mean squared error (MSE), Mean absolute error (MAE) may be used for the evaluation of the splitting process

These are the most critical parameters that should be tuned essentially in order to achieve the most satisfactory results. The RandomizedSearchCV tool in Python is the best choice to find the best parameters that lead to higher accuracy. This tool is explained in more detail in the different dedicated sections.

Logistic Regression hyperparameters (Classification)

Logistic regression as a binary classifier also has several parameters that may require a tuning process.

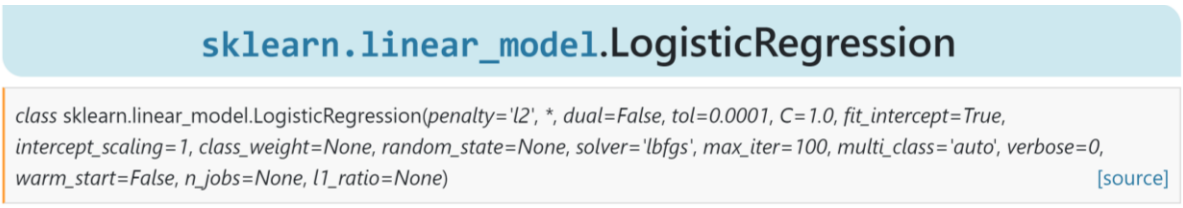


FIGURE 0.3 – LOGISITC REGRESSION HYPERPARAMETERS (HTTPS://SCIKIT-LEARN.ORG/STABLE/MODULES/GENERATED/SKLEARN.LINEAR_MODEL.LOGISTICREGRESSION.HTML)

Similar to RF, most of the parameters were kept unchanged during the process. However, some of them have been tuned, which are as follows:

- ***penalty (default = l1):*** This parameter indicates the type of penalty that is taken into account during the process. Different values can be defined for penalty parameters such as l1, l2, elasticnet, and none. Increasing the number of variables would cause complexity during the modeling process and probably would lead to overfitting. Regularization is a well-known technique that can be employed to prevent overfitting. Scikit-learn provides LASSO (l1), Ridge (l2), and Elasticnet (l1 and l2) regression to overcome overfitting. Due to the increasing intricacy of the model, the coefficient`s size escalates significantly, and applying the penalty term to the magnitude of the coefficients would handle this challenge (Swamynathan, 2019). The LASSO (l1) regression tries to minimize the coefficients of those variables that have a minor impact on the model. The Ridge regression (Tikhonov or l2), on the other hand, guides coefficients to be close to zero as much as possible but not zero itself. This penalty can be used when many variables add insignificant values to the model's accuracy individually but enhance efficiency and accuracy overall. Therefore, these variables cannot be excluded from the analysis (James et al., 2013; Swamynathan, 2019). The Elasticnet regression is a model that combines both LASSO and Ridge regression to find the best values for corresponding coefficients for different input variables.

- ***solver (default = "lbfgs"):*** This parameter is typically utilized in optimization problems. The algorithms that may be used for this parameter are sag, saga, lbfgs, and newton-cg.

sag: this method employs Stochastic Average Gradient descent. Speed-wise, sag is faster than other solvers and is recommended to be used for large datasets (a large number of data points and a large number of variables).

saga: is a variant of the sag algorithm. It only supports elasticnet regression penalty terms.

lbfgs: this is an optimization algorithm and is related to quasi-Newton, and Broyden-Fletcher-Goldfarb-Sahnno. This algorithm is typically used for small datasets.

Some of the solvers mentioned above merely work with either l1 penalty or l2 penalty. Therefore, this point should be taken into account during the modeling process.

XGBOOST (Classification and Regression)

Extreme Gradient Boosting (XGBOOST) is another state-of-the-art algorithm explained in detail in the previous chapter. Like other models, XGBOOST also has a bundle of hyperparameters, some of which should be explained in more detail. Following are the most influential parameters that should be considered during the tuning process. It should be noted that other parameters may be adjusted based on the accuracy and the outputs that practitioners are seeking.

- ***min_child_weight (default = 1):*** The minimum sum of weights related to all instances which is required in a child. If the step of partitioning step leads to a node with the sum weight of less than min_child_weight, then the partitioning will stop. The value for this parameter could be 0 to infinity.
- ***Booster (default = gbtree):*** This parameter defines the booster used for the partitioning process. It could be gbtree, gblinear, or dart.
- ***eta (default = 0.3):*** This parameter is the learning rate (step size shrinkage) used during the update of each booster, and it helps prevent overfitting. The value can be in the range of 0 to 1.
- ***Lambda (default =1):*** This is Ridge term (l2 regularization), with a default value of 1
- ***alpha (default = 0):*** This is parameter is LASSO term (l1 regularization). It should be noted that increasing both lambda and alpha would make the model more conservative.
- ***gamma (default = 0):*** This parameter defines the minimum loss reduction in order to make a subsequent partition on a leaf node. Same as lambda and alpha, increasing gamma would also make the model more conservative.

It should be mentioned that finding the best parameters for XGBOOST is time-consuming and should be done meticulously.

Artificial Neural Networks (Multi-Layer Perceptron)

The underlying concepts of ANN were explained in the previous chapter in precise detail. In addition, the Multi-Layer Perceptron (MLP) algorithm from Scikit-learn has been employed in this study. The given figures provided from Scikit-learn documentation indicate all hyperparameters that can be tuned for the MLP algorithm (Figure 0.4; Figure 0.5). In the following section, some of the most important parameters are briefly explained.

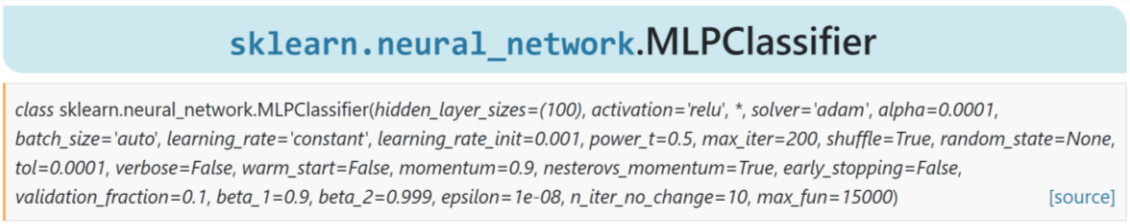


FIGURE 0.4 – MULTI-LAYER PERCEPTRON CLASSIFIER (HTTPS://SCIKIT-LEARN.ORG/STABLE/MODULES/GENERATED/SKLEARN.NEURAL_NETWORK.MLPCLASSIFIER.HTML)

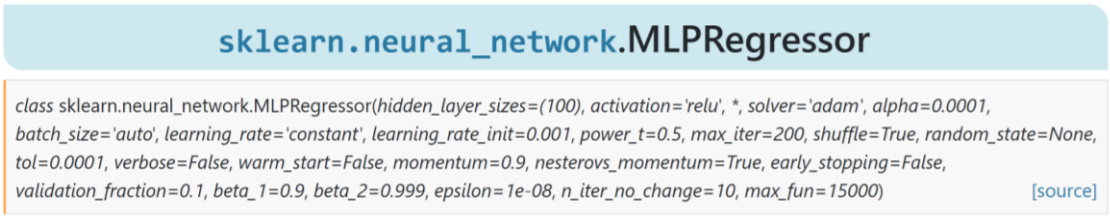


FIGURE 0.5 - MULTI-LAYER PERCEPTRON REGRESSOR (HTTPS://SCIKIT-LEARN.ORG/STABLE/MODULES/GENERATED/SKLEARN.NEURAL_NETWORK.MLPREGRESSOR.HTML)

- ***hidden_layer_sizes: (default = (100,)):*** Represents the number of neurons in ith layer of the MLP. For example, (25,15,) defines a model with two hidden layers with corresponding neurons, 25 and 15 for the first and second layers, respectively.
- ***activation (default = "relu"):*** activation parameter determines the activation function discussed before. These functions could be *sigmoid*, *tanh*, and *relu*. However, there are other activation functions, but only these three have been used during the tuning process.

Sigmoid *is a mathematical function with an output range between 0* and *1 (Figure* 0.6*)(Verdhan, 2020). This function is usually used for binary classification in the output layer. However, it can also be used within the hidden layers.*

$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{e^{x}+1}$$

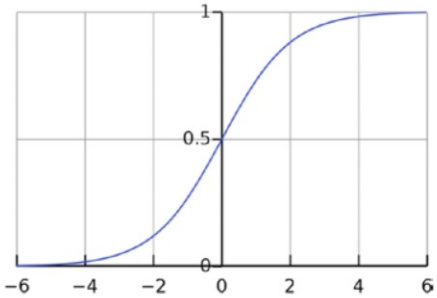


FIGURE 0.6 – SIGMOID ACTIVATION FUNCTION (VERDHAN, 2020)

tanh *or Tangent hyperbolic is an altered version of the sigmoid with a range between -1 and +1. This method is 0 based and is usually used for hidden layers* (Verdhan, 2020).

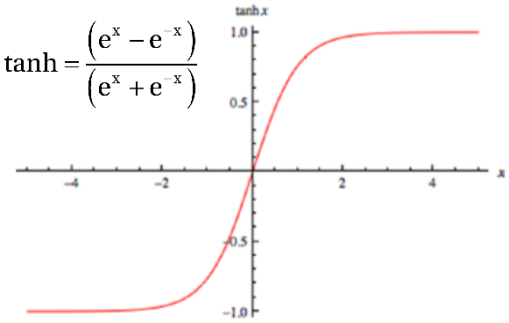


FIGURE 0.7 - TANH FUNCTION (VERDHAN, 2020)

relu *or Rectified Linear Unit is probably the most renowned activation function and is known for its simplicity. F(X) = max(x,0): this function provides the output of x for x>0, otherwise, 0 would be the output. The simplicity of ReLU makes it very straightforward and fast to train, and it is usually used for hidden layers.*

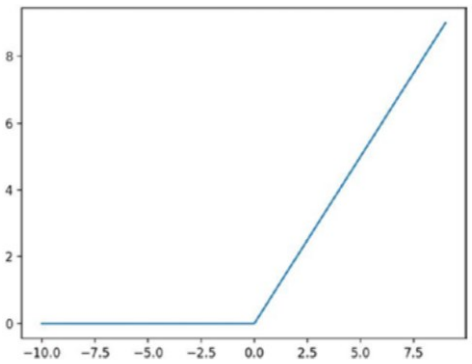


FIGURE 0.8 – RELU FUNCTION (VERDHAN, 2020)

Either of these functions can be chosen based on the results of cross-validation and evaluation metrics.

- ***Solver (default = "adam"):*** This parameter is used for optimization. Adam, sgd, and lbfgs are the algorithms that can be used for optimization.

Sgd *stands for stochastic gradient descent, and adam is another method based on sgd that other developers have introduced. Adam works well with large datasets in terms of validation and training time.*

lbfgs *is a method of optimization from the family of quasi-Newton*

- ***alpha (default = 0.0001):*** This parameter is l2 penalty and is used for optimization
- ***momentum (default = 0.9):*** This parameter is used for updating the gradient descent when tries to minimize the cost function.

It should be noted that hyperparameters for both classification and regression neural networks are the same and can be found using RandomizedSearchCV or GridSearchCV approaches.

Elasticnet Regression

Elasticnet regression is a combination of Ridge regression (ℓ2) and LASSO regression (ℓ1). This combination helps the model learn where few of the coefficients` weights are zero (LASSO) while keeping the settings of Ridge regression. In addition, this model would be helpful where there are, for instance, two highly correlated variables. In this case, LASSO would probably pick one of these attributes, whereas Elasticnet may select both variables (James et al., 2013; Rebala et al., 2019; Swamynathan, 2019; Scikit-learn Documentation). Figure 0.9 from Scikit-learn documentation shows all hyperparameters for this algorithm. Also, the most critical parameters of Elasticnet are as follow:

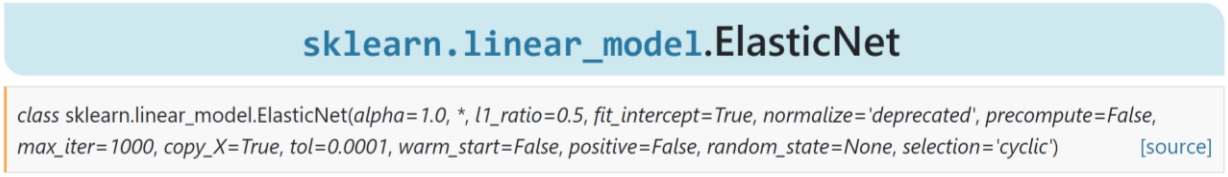


FIGURE 0.9 – ELASTICNET HYPERPARAMETERS (HTTPS://SCIKIT-LEARN.ORG/STABLE/MODULES/GENERATED/SKLEARN.LINEAR_MODEL.ELASTICNET.HTML)

- ***alpha (default = 1):*** Constant that multiplies the penalties. According to Scikit-learn documentation, if this value is considered to be 0, it would be equivalent to an ordinary least square, which Linear Regression solves.
- ***l1_ratio (default = 1):*** This parameter incorporate two penalty terms, ℓ1 and ℓ2. The range of this parameter is between 0 and 1. If 0, then the penalty would be l1, and if it equals 1, the penalty would be l1. For the range between 0 and 1, the combination of both penalties is used.
- ***max_iter (default=1000):*** Th maximum number of iterations could be done during the modeling process.

K-means Clustering

Cluster analysis is reported to have been the most popular unsupervised learning approach (Verdhan, 2020). This method typically partitions a dataset based on the similarity among the data points. In the water domain, this method can be applied for creating homogenous groups of pipes. This homogeneity can be based on different attributes such as material, diameter, length, and other available features. The clustering method was first introduced in the 1930s in the area of anthropology and psychology (Swamynathan, 2019).

K-Means is one of the most well-known approaches among clustering methods. This algorithm is an exquisite and straightforward method for partitioning a dataset into *K* discrete and un-overlapping clusters (James et al., 2013). In order to apply this method, first, the proper number of K should be determined; then K-means algorithm will go through the dataset and allocate each cluster to each data point. The provided figure indicates the clustering approach with the K-means method based on different values of K (Figure 0.10).

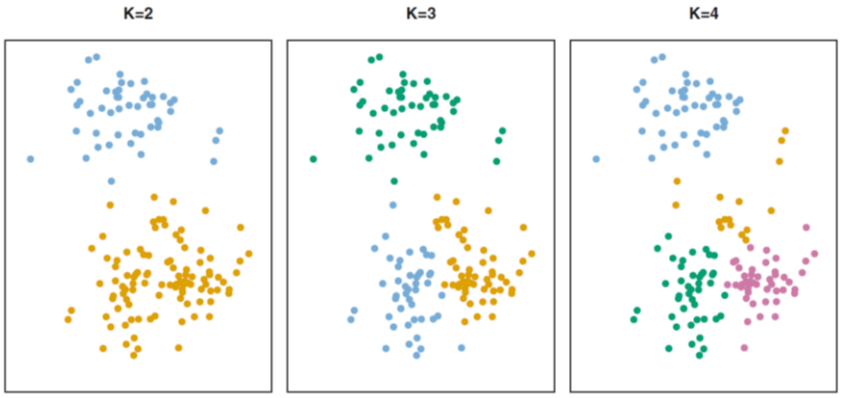


FIGURE 0.10 – K-MEANS CLUSTERING METHOD BASED ON DIFFERENT K(JAMES ET AL., 2013)

Following are the main two steps performed during the application of K-means clustering:

1- The number of K is defined. The K is the number of centroids used in K-means clustering. Then each data point is assigned to the nearest cluster. The centroid is the mathematical mean position of all data points.

2- In the next step, the centroid is recalculated based on the average coordinates of all the data points. It should be mentioned that K-Means is created based on Euclidean distance.

$$Euclidean\ Distance = d = \sqrt{\sum_{i=1}^{N}(X_i - Y_i)^2}$$

FIGURE 0.11 – EUCLIDEAN DISTANCE EQUATION (SWAMYNATHAN, 2019)

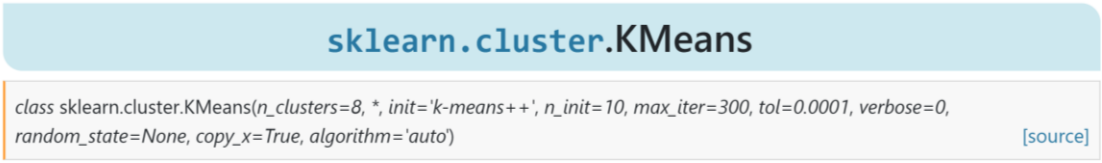


FIGURE 0.12 – K-MEANS CLUSTERING PARAMETRS (HTTPS://SCIKIT-LEARN.ORG/STABLE/MODULES/GENERATED/SKLEARN.CLUSTER.KMEANS.HTML)

The only parameter that has been adjusted in this study is the number of K (or n_clusters).

Classification and regression evaluation metrics

The main objective of developing a machine learning model is to predict the future (Verdhan, 2020). Hence, the most crucial step before deploying final models is the evaluation to find the best predictive model. Furthermore, evaluating the models helps specialists to better opt for the most efficient algorithm. It should be mentioned that this step typically varies between classification and regression models as each method has different evaluation metrics. The following sections provide more information regarding each metric and the way that they are calculated.

Confusion matrix (Accuracy, Precision, Recall, and f-score)

Confusion Matrix is a specifically designed table that is utilized to evaluate any classification algorithms (Swamynathan, 2019). This matrix is one of the most well-known approaches that can be used for either a binary classification or multiclass classification and is also represented as a two by two table for binary classifications (Verdhan, 2020). This method is an appropriate way to measure the efficiency of a machine learning model. From this matrix, various metrics can be extracted to evaluate the model performance. These metrics, for instance, could be accuracy, precision, recall, and f1-score, which are explained in the following sections. Figure 0.13 gives more insight regarding the confusion matrix and the values included in this table.

		Predicted	
		Negative (False)	Positive (True)
Actual	Negative (False)	True Negative (TN)	False Positive (FP)
	Positive (True)	False Negative (FN)	True Positive (TP)

FIGURE 0.13 – BASE CONFUSION MATRIX FOR A BINARY CLASSIFIER

Before going into more detail about the evaluation metrics, the values within the matrix should be clearly explained.

- ***True Negative or TN:*** This value is an indicator for Actual FALSE observations that are predicted correctly as False or Negative
- ***False Positive or FP:*** This value is an indicator for Actual FALSE observations that are mispredicted as True or Positive
- ***False Negative or FN:*** This value is an indicator for Actual TRUE observations that are mispredicted as False or Negative
- ***True Positive or TP:*** This value is an indicator for Actual TRUE observations that are predicted correctly as True or Positive

An acceptable model should have a relatively higher TN and TP than FN and FP (Verdhan, 2020). Furthermore, from the mentioned terminologies above, several metrics can be extracted. Following are the most important metrics for evaluation of a classification model:

- ***Accuracy:*** This metric indicates that how many or what percentage of predictions are made correctly. (when considering imbalanced data, this metric is not an appropriate choice)
- ***Precision:*** This metric shows that what percentage of positive predictions is correct
- ***Recall (Sensitivity or True-positive rate):*** This indicator shows what percentage of positive observations are caught by the model.
- ***F1 Score:*** F1 is a harmonic mean of Precision and Recall. For any imbalanced dataset, this is the best choice to evaluate the model. (Swamynathan, 2019; Verdhan, 2020)

- ***False-positive rate:*** This metric is an indicator that shows what percentage of actual False is predicted as True.

- ***AUC/ROC:*** Receiving operating characteristics curve (ROC) is employed for comparing different predictive models. This metric is a plot between True Positive Rate (TPR) and False Positive Rate (FPR). AUC, or the area under the ROC curve, is a measure that shows the goodness of the fit. This metric can be used as a final evaluation step where the result of classifiers are significantly close (Verdhan, 2020).

The given table provides more information about all metrics that can be calculated based on the confusion matrix (TABLE 0.1).

TABLE 0.1~~0.1~~ – CLASSIFICATION PERFORMANCE METRICS (SWAMYNATHAN, 2019)

Metric	Description	Formula
Accuracy	What % of predictions was correct?	(TP+TN)/(TP+TN+FP+FN)
Misclassification Rate	What % of prediction is wrong?	(FP+FN)/(TP+TN+FP+FN)
True Positive Rate OR SensitivityORRecall (completeness)	What % of positive cases did model catch?	TP/(FN+TP)
False Positive Rate	What % of No was predicted as Yes?	FP/(FP+TN)
Specificity	What % of No was predicted as No?	TN/(TN+FP)
Precision (exactness)	What % of positive predictions was correct?	TP/(TP+FP)
F1 score	Weighted average of precision and recall	2∗((precision ∗ recall) / (precision + recall))

MAE, MSE, RMSE, R-Squared

The difference between predicted values and actual conservations is called error which is made during the prediction step. This error should be minimized to achieve the best model. There are several approaches to evaluate the robustness of a regression model based on the error produced by the predictive models. These approaches are briefly explained in the following sections.

- ***Mean Absolute Error (MAE):*** From the name can be inferred that this metric is the average of absolute differences between predictions and actual values. The model tries to minimize this value

$$MAE = \frac{\sum(|\hat{y}_i - y_i|)}{n} \quad (12)$$

Where:

y_i is the actual value; $\hat{y}_i$ is the predicted value; and n is the number of observations

- ***Mean Squared Error (MSE):*** This metric indicates the average of the squared error.

$$MSE = \frac{\sum(|\hat{y}_i - y_i|)^2}{n} \quad (13)$$

- ***Root Mean Squared Error (RMSE):*** The square root of MSE is known as RMSE. This metric should be used to find out how close the predicted values and actual values are.

$$RMSE = \sqrt{MSE} \quad (14)$$

- ***R-Squared or R^2:*** It shows the total portion of the variance in the dependent variable, and it is one of the most popular metrics for evaluating a regression model. This value falls between 0 and 1 (Swamynathan, 2019). The closer this value is to 1, the more accurate the regression model is

$$SST\ (Sum\ of\ square\ total) = (y_i - \bar{y})^2 \quad (15)$$

$$SSR\ (Sum\ of\ sqaure\ residuals) = (\hat{y}_i - \bar{y})^2 \quad (16)$$

$$R - Squared = \frac{\sum SSR}{\sum SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \quad (17)$$

Where:

$\bar{y}$ is the mean of dependent variables

It should be noted that the interpretation of these metrics significantly relies on the domain that the model is created based on its dataset.

Validation method (n-fold cross-validation) – Train, Validate, and test

(This section is provided based on information on Scikit-learn documentation site)

It is an unjustifiable action if the same dataset is used for both training and testing the model. This type of model usually does not perform well on an unseen dataset, which is called overfitting. In order to prevent this phenomenon, a specific part of the dataset is held out as a test set. Learning the pattern on the training dataset, the predictive model then tests the model on the test samples. For instance, an algorithm can be trained on 80% of the dataset and tested on the remaining 20%. Nonetheless, a model created based on training and test set is still prone to overfitting. One way to address this challenge is to put another part of a dataset as a validation set. The model is then learned based on the training set and validated with the validation set. Should the results be satisfactory, then the model can be evaluated on the test set.

An important downside of partitioning a dataset into three parts is that the number of instances declines substantially, which can be used for the learning process. A well-known solution to this issue is a method so-called cross-validation. In this method still, a test set should be put aside for the final evaluation. However, the validation set is no longer required when employing the cross-validation method. In the method called k-fold cross-validation, the training set is divided into k parts. Then a model is learned using k – 1 folds from the training set. The remaining fold of data is validated based on the output model. The model is trained and validated k times, and the result is eventually the average of the numbers calculated in the recurring loop. Financially-wise, this method could be expensive even though it employs as many data points as possible during the training process. Figure 0.14**Error! Reference source not found.** indicates the concept of cross-validation, which is extracted from Scikit-learn documentation.

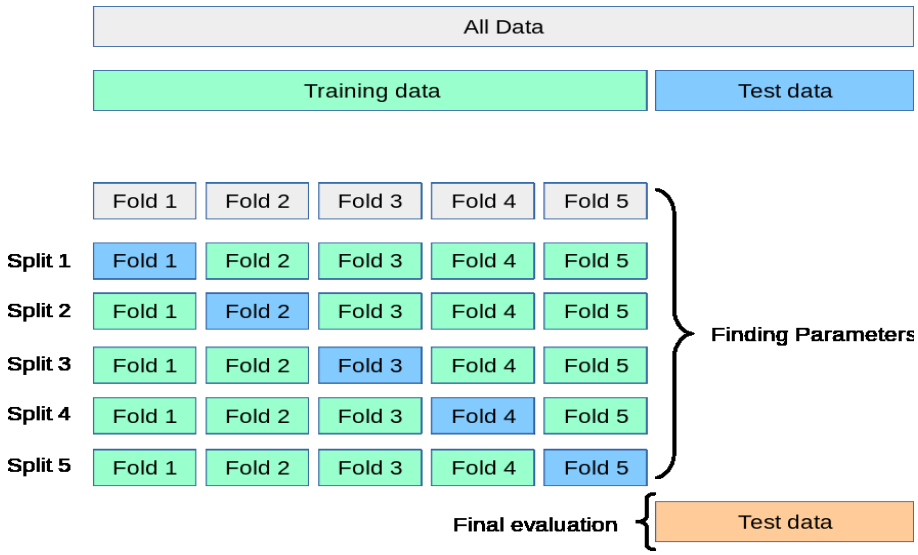


FIGURE 0.14 – CROSS-VALIDATION SCHEME BASED ON SCIKIT-LEARN DOCUMENTATION

GridsearchCV and RandomizedsearchCV

- ***GridsearchCV:*** For a specific machine learning model, it is possible to define a list of hyperparameters that are worth trying. Using GridsearchCV provided by Scikit-learn, the model is built based on all possible combinations of defined parameters (Swamynathan, 2019). The best combination is selected based on the previously mentioned cross-validation method. However, the GridsearchCV method is computationally expensive when the number of parameters increases. For example, consider 5-fold cross-validation for three parameters and each parameter with six predefined values. The number of combinations for this Gridsearch would be 3645, one of which should be selected as the best combination. Given figure indicates all parameters that GridsearchCV can use. The most critical parameters are the ***estimator*** (base model, such as the RandomForestClassifier), ***scoring*** (can be f1 score for classifications), ***CV*** (number of folds for cross-validation, and ***param_grid*** (dictionary including desired parameters).

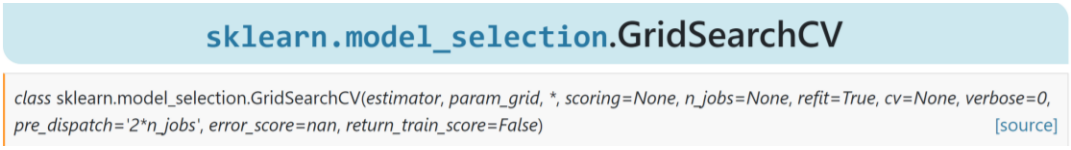


FIGURE 0.15 – GRIDSEARCHCV HYPERPARAMETERS (HTTPS://SCIKIT-LEARN.ORG/STABLE/MODULES/GENERATED/SKLEARN.MODEL_SELECTION.GRIDSEARCHCV.HTML)

- ***RandomizedSearchCV:*** As from the name can be inferred, numerical parameters can be defined as a range, unlike GridsearchCV, so that the algorithm would be able to search for the best parameters randomly. The number of iteration and combinations that can be used is readily defined. However, the iteration number should be adjusted carefully as missing the best parameters in this method is very likely to happen. Almost all parameters are similar to that of GridSearchCV except for a few parameters. ***param_distribution*** (dictionary including desired parameters and ranges), and ***n_iter*** (number of parameters sampled) are among these parameters. The given figure shows the parameters for Random Search, based on Scikit-learn documentation.

sklearn.model_selection.RandomizedSearchCV

class sklearn.model_selection.RandomizedSearchCV(*estimator, param_distributions, *, n_iter=10, scoring=None, n_jobs=None, refit=True, cv=None, verbose=0, pre_dispatch='2*n_jobs', random_state=None, error_score=nan, return_train_score=False*) [source]

FIGURE 0.16 - RANDOMIZEDSEARCHCV HYPERPARAMETERS (HTTPS://SCIKIT-LEARN.ORG/STABLE/MODULES/GENERATED/SKLEARN.MODEL_SELECTION.RANDOMIZEDSEARCHCV.HTML)

Based on the number of data points and the time required for the modeling process, one of the mentioned methods is used for different utilities in this study.

An Endeavour for handling imbalanced datasets using Smote method

As previously discussed in chapter 2 of this study, the existence of an imbalanced dataset is an inevitable part of data analytic and machine learning. Nonetheless, some oversampling and under sampling methods, such as Synthetic Minority Oversampling Technique (SMOTE), can be used to cope with imbalanced datasets. The algorithm performs the oversampling technique to rebalance the base training set. Instead of simply duplicating the minority class data points, the principle idea of SMOTE is to produce artificial instances (Fernandez et al., 2018). The new sample is produced based on the similarity between several minority class samples with a determined distance. This algorithm is created based on the feature space instead of the data space, which means that the algorithm is learned and created based on the features` values. For instance, assume $X_i$ is a minority class`s sample, a base sample for creating artificial samples. Based on the distance metric (Euclidean Distance), several nearest neighbors (X_{i1} to X_{i4}) are selected from the training part of the dataset. Eventually, an interpolation is performed to create new synthetic samples (r_1 to r_4) based on the nearest neighbors (Figure 0.17).

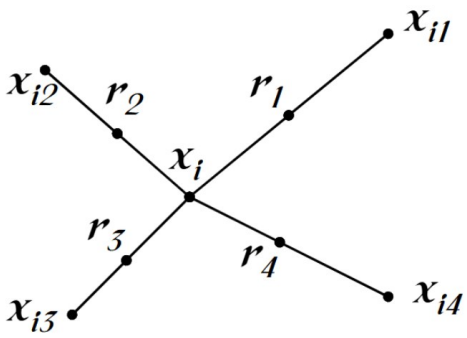


FIGURE 0.17 – CREATING SYNTHETIC INSTANCES IN THE SMOTE ALGORITHM (FERNANDEZ ET AL., 2018)

Using this approach may help to overcome the challenges caused by the imbalanced dataset. Results regarding the impact of SMOTE on the accuracy of the models are provided in chapter 5 of this study.

Normalization

In some cases, the unit of provided input variables may vary; therefore, the modeling results may be skewed toward those variables with higher values (Swamynathan, 2019). For instance, in this study, age and length have different ranges of values, four years and 1000 m, respectively. Therefore, transforming these values to the same range would overcome the skewing issue.

The Standardization (z-score) method has been employed for this study, making the mean value 0 and the standard deviation 1 (Swamynathan, 2019). Given equation shows that how a value can be standardized based on this assumption.

$$Standardized\ X = \frac{(\ x - mean\ )}{(standard\ deviation)} \tag{18}$$

StandardScaler function from Scikit-Learn was used for this step of the study. Therefore, all values were transformed to the same range for the models that required standardization. One of the models, for instance, is ANN that requires normalization before the learning process.

Dummy Variables

In order to be able to run algorithms in python, the categorical attributes, for instance, material, should have been transformed to a new numerical format. Hence, wherever the value of the categorical attribute is present, one is given to the cell; otherwise, 0 is given to that cell (Swamynathan, 2019; Verdhan, 2020). Thus, for instance, if the cast iron pipe exists for a specific data point, then that cell is given 1, else it would be 0. For this purpose, the 'get_dummies' function from python has been used.

Overfitting and Underfitting Control

When creating a predictive model, overfitting and underfitting are common challenges (Verdhan, 2020), and they are often called bias-variance tradeoffs. Underfitting happens when the model cannot learn a desirable pattern while being trained, and the model is weak. On the other hand, overfitting is the case when a model learns patterns from the dataset more

accurately, although the result for the test set does not show satisfactory performance. The given figures provide the concept of overfitting and underfitting (Figure 0.18; Figure 0.19).

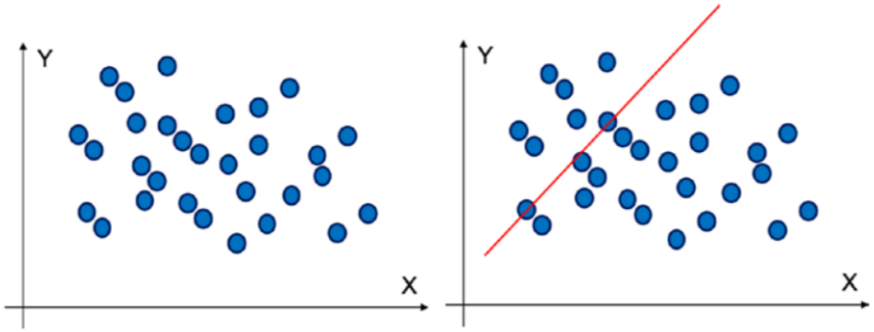


FIGURE 0.18 – UNDERFITTING WHEN A VERY SIMPLE MODEL HAS BEEN PRODUCED (VERDHAN, 2020)

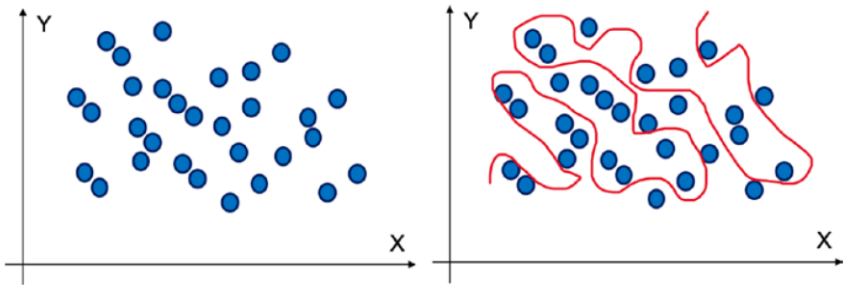


FIGURE 0.19 – OVERFITTING WHERE A VERY INTRICATE MODEL HAS BEEN PRODUCED (VERDHAN, 2020)

There are different approaches to find out whether there is either overfitting or underfitting for the created models. One of these methods is plotting the F1-Score for training and test sets based on specific criteria.

The random forest as an example was chosen to make this step more clear to perceive. As previously mentioned, there are several hyperparameters for random forest algorithm. However, maximum depth is the one that may control overfitting and underfitting conditions. Therefore, when the model is created based on different parameters, it is also tested based on the different values for maximum depth. Therefore, this model was analyzed in two steps: based on the F1-Score and the improvement to detect correct values (improve misclassification values).

The given figure is created based on the random forest algorithm, which plots F1-Score as the most critical metric in this study versus maximum depth defined for the model (Figure 0.20). It is evident that, although insignificant, overfitting does exist in this case. For example, after a depth of 8, the result for the test set worsened compared to the training set. Therefore, this

graph shows that the number for maximum depth should be increased carefully in order to prevent overfitting.

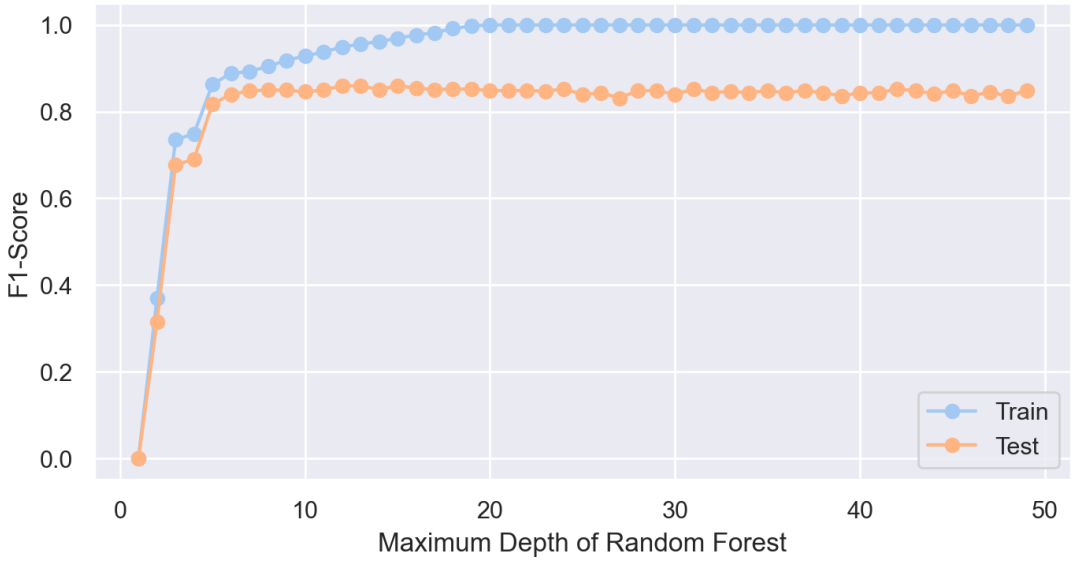


FIGURE 0.20 – TEST RANDOM FOREST MODEL FOR OVERFITTING AND UNDERFITTING, BASED ON THE F1-SCORE (MARKHAM)

Furthermore, the random forest model also was analyzed based on the number of misclassification values. The given line graph plots number of misclassified values and the maximum depth defined for the random forest algorithm (Figure 0.21).

An improvement can be seen when the depth of trees increases. This enhancement starts with over 120 misclassified data points for a depth of one, and the number of misclassified samples declines abruptly, as depth surges from 1 to 7, reaching a minimum of around 35 misclassified instances. However, no significant improvement can be noticed from the depth of 7 to above within the graph. This indicates that there is no need to use a maximum depth of more than 7 in this study. Thus, other hyperparameters may be tuned in order to decrease the number of misclassified samples.

This test only considers maximum depth, which is insufficient to ensure that the model over fits or under fits the dataset. Therefore, all influential hyperparameters should be evaluated.

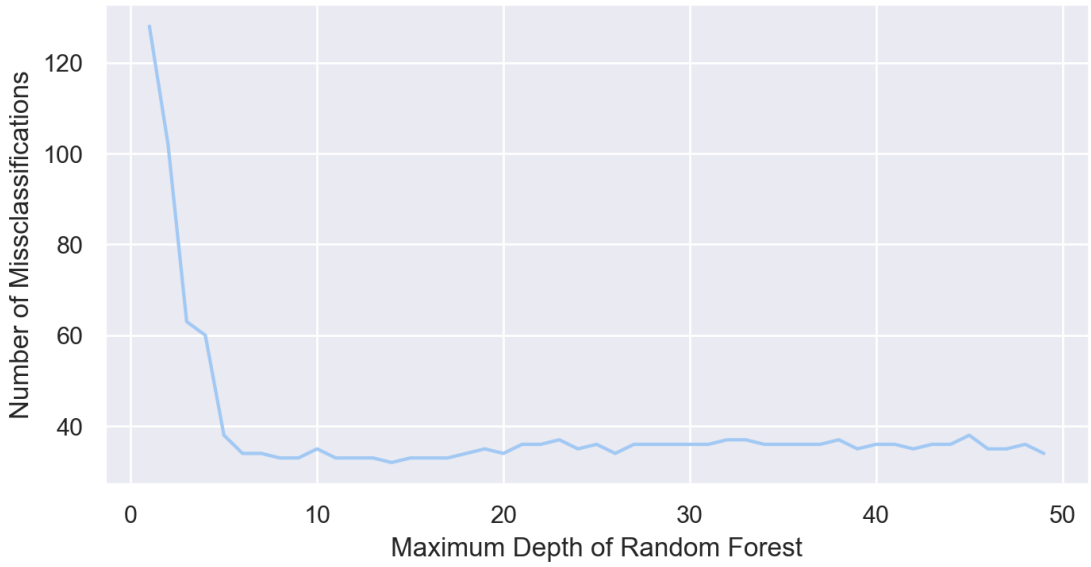


FIGURE 0.21 - TEST RANDOM FOREST MODEL FOR OVERFITTING AND UNDERFITTING, BASED ON THE NUMBER OF MISCLASSIFIED SAMPLES (MARKHAM)

APPENDIX C – CLASSIFICATION RESULTS (ALL CITIES IN DETAIL)

Saskatoon

In this section, results related to the city of Saskatoon are provided. After cleaning and preparing the classification dataset, 32,306 pipes remained, including different attributes; diameter, material, joint type, length, lining material, lining status, lining age, age, and more importantly, target (dependent variable).

PVC pipe is the most frequent type in terms of material, which accounts for 56.56% of non-broken pipes, followed by asbestos cement and cast iron with 28.11% and 12.39% contribution. However, cast-iron makes up 48.77% among broken pipes, slightly more than asbestos cement material with a 46.28% contribution. It is worth mentioning that other materials within the Saskatoon network have a small proportion. The lack of sufficient information for these types of pipes would decrease the accuracy of the final predictive model. Nonetheless, all materials have been used for the modeling process in the 1st step. Furthermore, the most frequent materials have also been analyzed separately, and related results are provided in the appendix part of this study. The following bar chart represents more information considering different materials within this classification dataset (Figure 0.1).

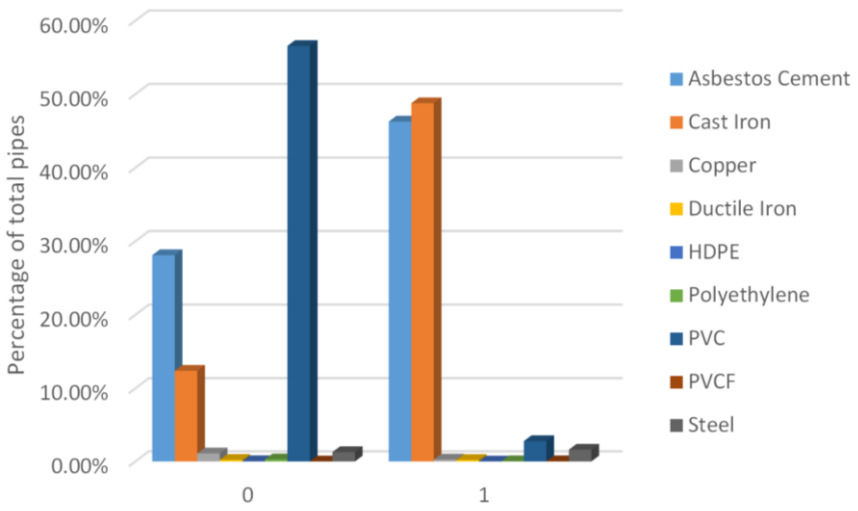


FIGURE 0.1 - PERCENTAGE OF EACH MATERIAL FOR BOTH CLASSES (0,1) BASED ON TOTAL PERCENTAGE OF EACH CLASS (SASKATOON)

As previously mentioned, there are 32,306 pipes within the network, including 28,635 non-broken pipes (class 0) and 3,671 broken pipes (class 1). Dataset is split into train, validation, and test set (train, validation = 80%, test = 20%). After the splitting process, there are 5,731 non-broken pipes and 731 broken pipes for the test set. Given is the confusion matrix which has been prepared based on the evaluation of the test (TABLE 0.1). Explicitly can be seen that the XGBOOST classifier was able to detect the highest number of broken pipes. However, random forest predicted non-broken pipes more accurately, with the number of 5,700.

TABLE 0.~~10.~~1 - CONFUSION MATRIX FOR ALL MATERIALS (SASKATOON)

Random Forest		**Predicted**		**XGBOOST**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*172*	*559*	**Actual**	**1**	*142*	*589*
	0	*5700*	*31*		**0**	*5686*	*45*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*242*	*489*		**1**	*153*	*578*
	0	*5658*	*73*	**Actual**	**0**	*5675*	*56*

0 = None-Broken
1 = Broken

From the extracted results, XGBOOST was found to be the best algorithm to detect a pattern in the dataset with accuracy and F1-score of 97% and 89%, respectively (TABLE 0.2). In order to find out how a homogenous group of pipes could affect the accuracy, other materials were analyzed separately. The result for cast iron pipes shows approximately a similar accuracy to that of all materials.

SMOTE has also been applied to the entire dataset, including all materials, to determine whether the overall accuracy can be improved. The results represented that SMOTE method decreased the model`s performance, showing that oversampling methods do not necessarily improve the evaluation scores. For instance, the F1-Score for XGBOOST declined from 89% to 83% in the SMOTE method, or the ANN model's result remained unchanged. The ANN algorithm with an 85% F1-score has the most satisfactory performance for SMOTE algorithm. Other results regarding different materials are provided in the Appendix section.

TABLE 0.2~~0.2~~ – CLASSIFICATION RESULTS (SASKATOON)

Algorithm	Accuracy			F1 - Score		
	AM	SMOTE	Cast Iron	AM	SMOTE	Cast Iron
Random Forest	*96%*	*95%*	*92%*	*85%*	*81%*	*86%*
XGBOOST	*97%*	*96%*	*94%*	*89%*	*83%*	*89%*
Logistic Regression	*95%*	*90%*	*86%*	*78%*	*66%*	*78%*
ANN	*97%*	*97%*	*92%*	*85%*	*85%*	*86%*

**** AM = All Materials,***
**** SMOTE = All metrials with SMOTE method***

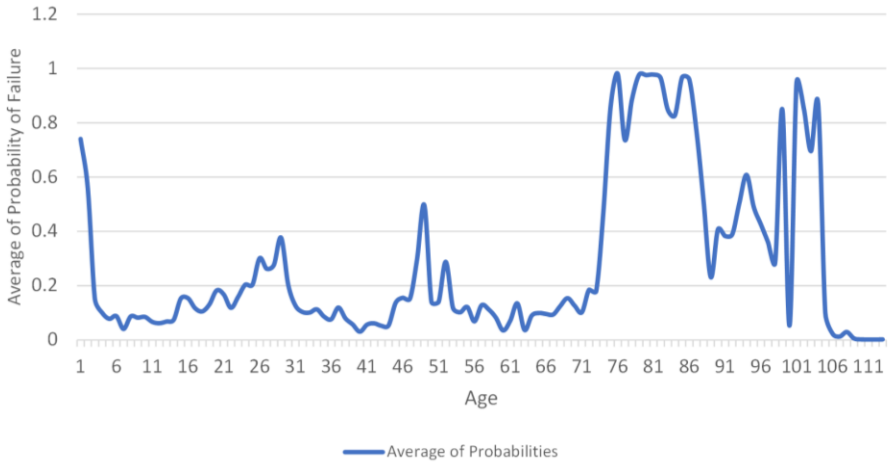


FIGURE 0.2 – AVERAGE OF PROBABILITY OF FAILURE BASED ON AGE (SASKATOON)

Winnipeg

The most comprehensive dataset has been provided by Winnipeg utility. After the cleaning process, this dataset includes 102,631 samples containing different variables such as material, diameter, length, coating material, age, and the target variable.

Similar to Saskatoon's, PVC pipes have the highest contribution to non-broken pipes, which is 67.85%. For broken pipes, cast iron accounts for almost 70.35% of the historical records, followed by asbestos cement pipes and ductile iron, with 22.19% and 4.29% in successive. However, again for the majority of materials, no adequate information is available. Therefore, the results for these specific materials are not as much reliable as, for instance, cast iron and asbestos cement pipes. Machine learning algorithms are typically sensitive to cases where enough information is not available since they cannot learn a pattern from the datasets efficiently and adequately. Below is the figure that shows the percentage of each material based on their classes (Figure 0.3).

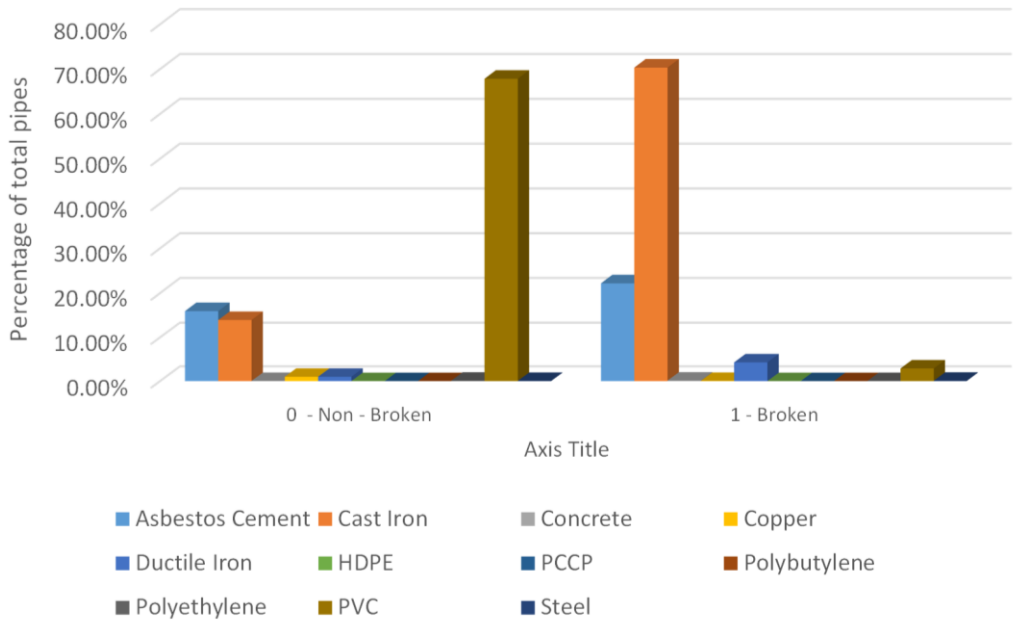


FIGURE 0.3 - PERCENTAGE OF EACH MATERIAL FOR BOTH CLASSES (0,1) BASED ON TOTAL PERCENTAGE OF EACH CLASS (WINNIPEG)

From the entire data points within the dataset, 94,498 pipes are non-broken, and 8,133 are broken. Thus, even though the dataset format may be considered imbalanced, many broken pipes can be employed in the learning process. As discussed before, 20% of these pipes belong to the test set, including 18,898 class 0 pipes and 1,629 class 1 pipes. Confusion for all materials is provided below, which compares the power of prediction between various models (TABLE 0.3). Again the XGBOOST algorithm showed better performance for detecting broken pipes

(class 1), and random forest with an infinitesimal difference could predict the maximum number of non-broken pipes.

Nonetheless, the XGBOOST classifier for all materials and cast iron pipes indicated a better performance with an F1-Score of 74% and 75%, respectively (TABLE 0.4). Like Saskatoon, the ANN was the best model for the SMOTE algorithm with an F1-Score of 73%, which remained unchanged for Winnipeg. F1-Score decreased from AM to SMOTE, indicating that SMOTE may not be used where there is sufficient information for the minority class.

TABLE 0.3~~0.3~~ - CONFUSION MATRIX FOR ALL MATERIALS (WINNIPEG)

Random Forest		**Predicted**		**XGBOOST**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*651*	*978*	**Actual**	**1**	*563*	*1066*
	0	*18786*	*112*		**0**	*18730*	*168*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*929*	*700*		**1**	*577*	*1052*
	0	*18644*	*254*	**Actual**	**0**	*18709*	*189*

0 = None-Broken
1 = Broken

TABLE 0.4~~0.4~~ - CLASSIFICATION RESULTS (WINNIPEG)

Algorithm	**Accuracy**			**F1 - Score**		
	AM	**SMOTE**	**Cast Iron**	**AM**	**SMOTE**	**Cast Iron**
Random Forest	*96%*	*92%*	*87%*	*72%*	*62%*	*74%*
XGBOOST	*96%*	*94%*	*87%*	*74%*	*68%*	*75%*
Logistic Regression	*94%*	*87%*	*77%*	*54%*	*52%*	*65%*
ANN	*96%*	*96%*	*86%*	*73%*	*73%*	*74%*

**** AM = All Materials,***
**** SMOTE = All metrials with SMOTE method***

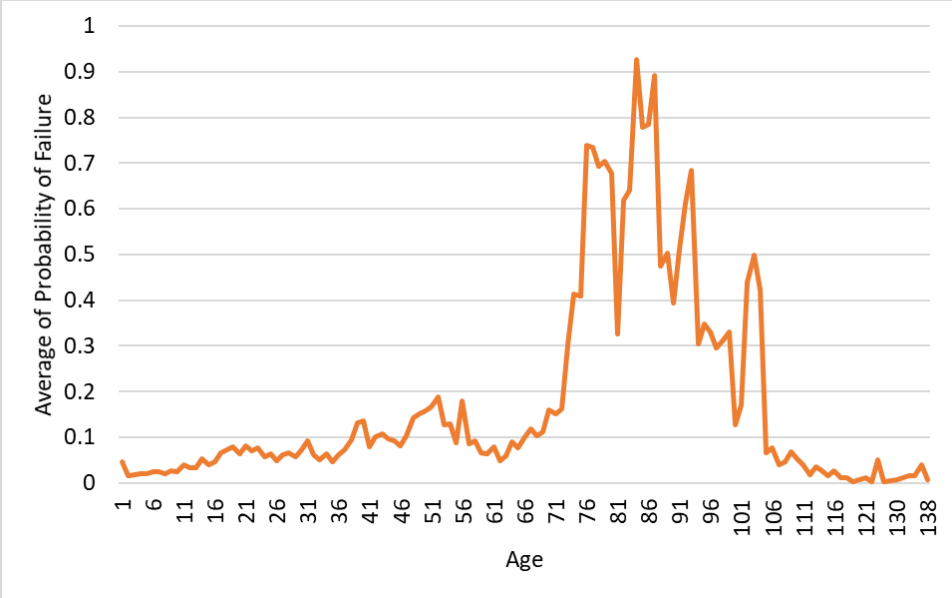


FIGURE 0.4 - AVERAGE OF PROBABILITY OF FAILURE BASED ON AGE (WINNIPEG)

Kitchener

In this section, results related to the city of Kitchener located in Ontario province are provided. This dataset consists of 14,568 segments and various input variables such as Material, LiningMaterial, Diameter, LiningStatus, Length, Age, LiningAge, and more importantly, the target attribute.

For non-broken pipes, PVC and DI have the highest frequency in the network, with 40% and 37%, in successive, followed by CI pipes with a 16.65% contribution. The PVCP and Concrete pipes are among the other materials with having a small proportion of the inventory records (Figure 0.5).

On the other hand, CI with 69.52% of the entire incidents is a predominant type of material for broken pipes. Ductile Iron is another material with a relatively significant number of recorded failures, 27.52%, followed by PVC with 2% involvement. Thus, cast Iron pipes seem to be in an extreme deterioration condition among most of the networks in this study.

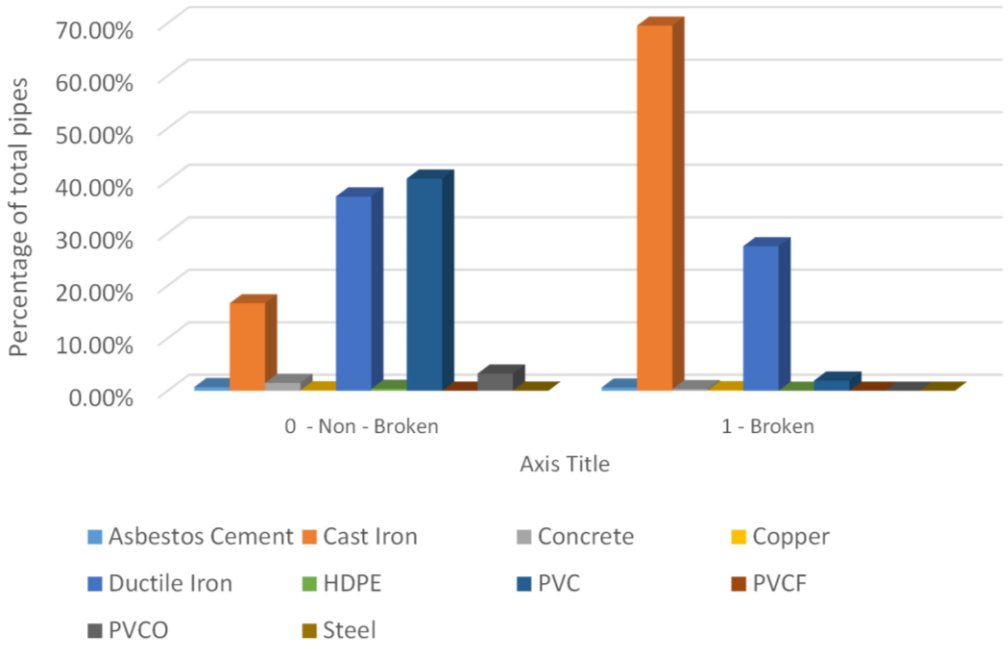


FIGURE 0.5 - PERCENTAGE OF EACH MATERIAL FOR BOTH CLASSES (0,1) BASED ON TOTAL PERCENTAGE OF EACH CLASS (KITCHENER)

For the Kitchener network, 13,587 pipes are non-broken, and 981 are collected as broken. Therefore, this dataset may be assumed to be an imbalanced type. With a 20% test size, the number of broken pipes is 210, and the non-broken pipe is 2,704.

TABLE 0.5~~0.5~~ - CONFUSION MATRIX FOR ALL MATERIALS (KITCHENER)

Random Forest		**Predicted**		**XGBOOST**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*109*	*101*	**Actual**	**1**	*103*	*107*
	0	*2701*	*3*		**0**	*2680*	*24*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	

		0	1			0	1
Actual	**1**	*117*	*93*		**1**	*100*	*110*
	0	*2692*	*12*	**Actual**	**0**	*2689*	*15*

0 = None-Broken, 1 = Broken

For the Kitchener dataset (AM), ANN showed the accuracy and F1-score of 96% and 66%, respectively. Thus, this algorithm is considered to be the best one for this network when considering all materials. However, when the dataset was partitioned based on different materials, Cast Iron showed a higher F1-score, although it showed a lower accuracy than all materials. Therefore, the XGBOOST is the best classifier for Cast Iron pipes in the Kitchener with an F1-score of 78%. SMOTE method was also utilized for having better accuracy. However, this method did not show a good performance.

Furthermore, the SMOTE method does not perform well with all kinds of imbalanced datasets. Hence, the results should always be compared to determine whether the oversampling method can improve the algorithm's performance. The given figure compares different models based on the accuracy and F1-score (TABLE 0.6).

TABLE 0.6~~0.6~~ - CLASSIFICATION RESULTS (KITCHENER)

Algorithm	**Accuracy**			**F1 - Score**		
	AM	**SMOTE**	**Cast Iron**	**AM**	**SMOTE**	**Cast Iron**
Random Forest	*96%*	*86%*	*91%*	*64%*	*46%*	*77%*
XGBOOST	*96%*	*90%*	*91%*	*63%*	*54%*	*78%*
Logistic Regression	*96%*	*83%*	*89%*	*59%*	*42%*	*73%*
ANN	*96%*	*96%*	*90%*	*66%*	*66%*	*77%*

**** AM = All Materials,***

**** SMOTE = All metrials with SMOTE method***

Markham

Kitchener is another utility analyzed in this study, with 10,786 pipes within the classification dataset. There are different input variables for this part of the study, including material, diameter, length, lining status, protection status, protection age, lining age, and finally, target variable.

Among non-broken pipes, PVC pipes are the predominant type of material with 79.51%, followed by ductile iron with a 12.37% contribution. Other materials account for small proportions of non-broken pipes. Furthermore, considering broken pipes, ductile iron with 54.98% is the primary type of material that experienced more failures than others. This material is followed by cast iron and PVC pipes with 32.95%, and 9.62% recorded failures, respectively. The given figure compares both class 0 and class 1 in terms of material frequency (Figure 0.6).

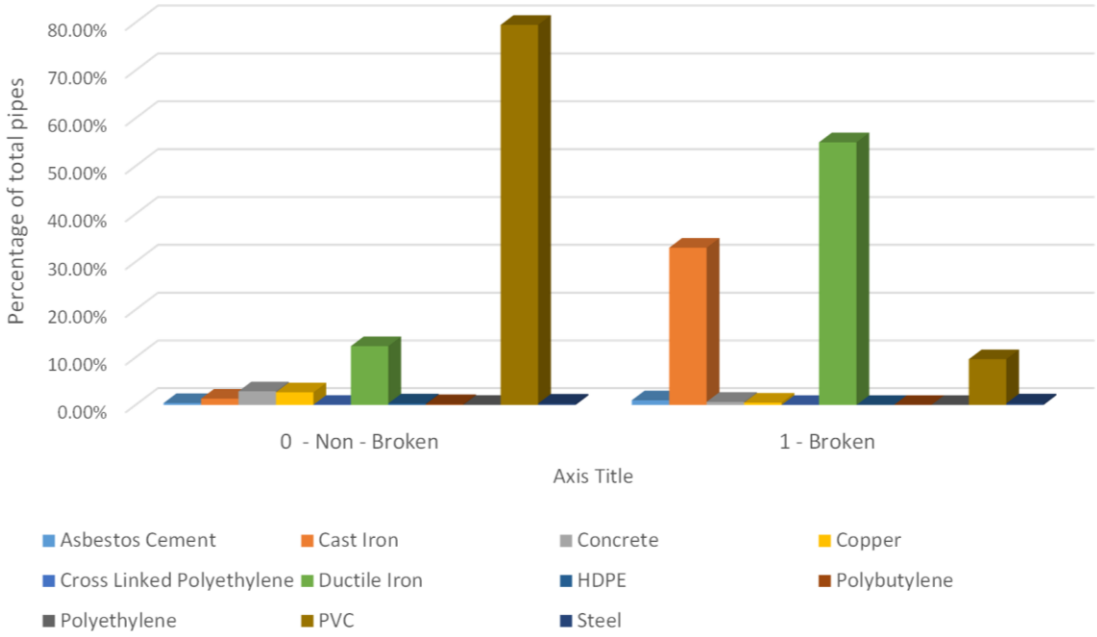


FIGURE 0.6 - PERCENTAGE OF EACH MATERIAL FOR BOTH CLASSES (0,1) BASED ON TOTAL PERCENTAGE OF EACH CLASS (MARKHAM)

The number of failures experienced a peak for pipes with the age of 10 to 30. From this, it can be inferred that younger pipes are experiencing most failures within the Markham network. However, it should be noted that the distribution of age may vary among different utilities based on the pre-defined framework for data collection. More importantly, most of the failures happened during the in-usage step of the Bathtub curve.

There are 10,786 segments within the classification dataset, among which 10,173 are non-broken pipes and 613 are broken pipes. From these numbers, it is clear that the type of dataset can be considered as imbalanced. Like other utilities, the confusion matrix is created based on a 20% test set, including 2,030 pipes in class 0 and 128 pipes in class 1. From the given table, it can be noticed that the accuracy scores for all groups are relatively similar, with 98% for all materials. However, when the models are compared with F1-Score, ANN showed a better performance with 86%. For all materials using SMOTE algorithm, ANN again indicated a better performance with an 86% F1-Score, showing that the performance of ANN usually stays unchanged between all categories as opposed to other algorithms, which shows decreasing accuracy when using SMOTE. Furthermore, logistic regression represented a better performance for CI pipes with 96% of F1-Score despite all models having approximately similar

scores for this group of analysis. The following table demonstrates the overall accuracy of these classifiers, and more information is provided in the appendix section of this study (TABLE 0.7; TABLE 0.8).

TABLE 0.70.7 - CONFUSION MATRIX FOR ALL MATERIALS (MARKHAM)

Random Forest		Predicted		XGBOOST		Predicted	
		0	1			0	1
Actual	1	*32*	*96*	Actual	1	*30*	*98*
	0	*2028*	*2*		0	*2026*	*4*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*36*	*92*		**1**	*26*	*102*
	0	*2024*	*6*	**Actual**	**0**	*2023*	*7*

0 = None-Broken
1 = Broken

TABLE 0.80.8 - CLASSIFICATION RESULTS (MARKHAM)

Algorithm	Accuracy			F1 - Score		
	AM	SMOTE	Cast Iron	AM	SMOTE	Cast Iron
Random Forest	*98%*	*98%*	*94%*	*84%*	*83%*	*95%*
XGBOOST	*98%*	*98%*	*94%*	*85%*	*80%*	*95%*
Logistic Regression	*98%*	*96%*	*96%*	*81%*	*70%*	*96%*
ANN	*98%*	*98%*	*93%*	*86%*	*86%*	*94%*

**** AM = All Materials,***

**** SMOTE = All metrials with SMOTE method***

Waterloo

Like other cities, followed by the data cleaning process, the classification dataset was prepared to calculate the probability of failure for the Waterloo network. This dataset consists of 7,532 pipes with different attributes such as diameter, material, length, lining status, lining material, lining age, age, and the target variable.

Various types of materials have been installed in Waterloo, a list of which can be seen in the given figure. PVC pipes account for almost 58.45% of the non-broken pipes, followed by CI and DI, which make up 26.13% and 14.80%, respectively. The CI pipes, once more, DI pipes are the most frequent type, with 81.71% and 15.28%, among total broken pipes.

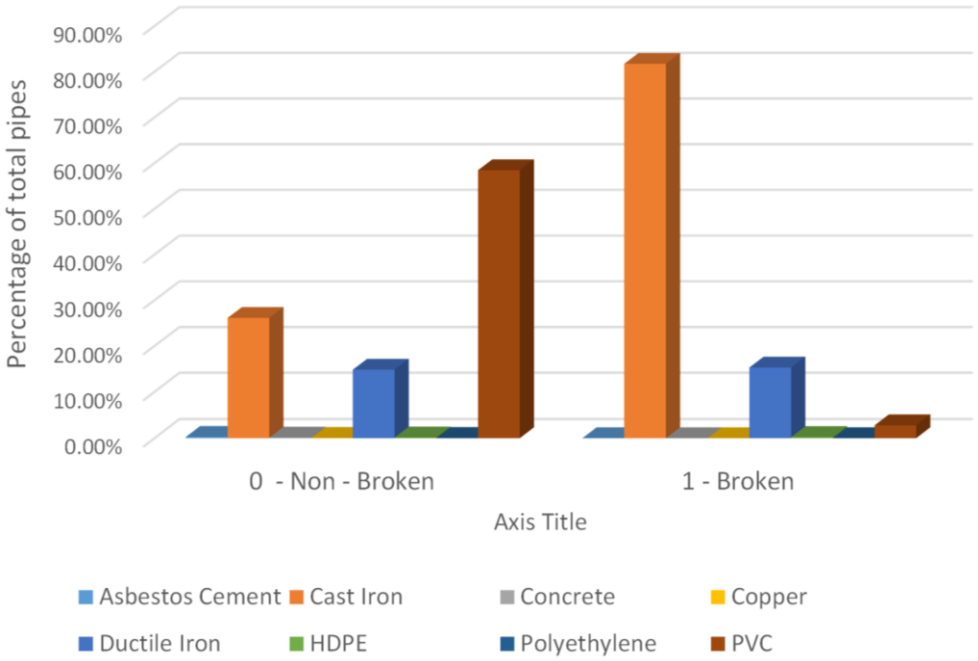


FIGURE 0.7 - PERCENTAGE OF EACH MATERIAL FOR BOTH CLASSES (0,1) BASED ON TOTAL PERCENTAGE OF EACH CLASS (WATERLOO)

As previously mentioned, the target variable includes two classes; class 0 and class 1. Among the entire pipes, 7,100 are related to class 0 or non-broken, and 432 pipes are related to class 1 or broken pipes. Thus, 20% of these pipes belong to the test set with 1,402 pipes for class 0 and 105 pipes for class 1. Therefore, the following confusion matrix was prepared from this test set, and XGBOOST represented a better performance in predicting broken pipes (TABLE 0.9).

TABLE 0.9~~0.9~~ - CONFUSION MATRIX FOR ALL MATERIALS (WATERLOO)

Random Forest		**Predicted**		**XGBOOST**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*72*	*33*	**Actual**	**1**	*61*	*44*
	0	*1397*	*5*		**0**	*1392*	*10*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*80*	*25*		**1**	*65*	*40*

	0	*1392*	*10*	**Actual**	**0**	*1391*	*11*

0 = None-Broken
1 = Broken

Due to the imbalanced format of the dataset, the predictive models were not able to show satisfactory performance. Logistic regression with an F1-score of 36% was the weakest classifier for AM and SMOTE categories. XGBOOST, however, indicated a better F1-Score among these algorithms. Accuracy and F1-Score for XGBOOST are 95% and 55% for all materials, and in the best case, F1-Score for cast iron pipes is 67%. Comparing the results between different categories indicates that partitioning pipe into homogenous groups may increase the performance of machine learning models. Results related to other models are provided in the table (TABLE 5.19), and further information is found in the appendix section.

TABLE 0.10~~0.10~~ - CLASSIFICATION RESULTS (WATERLOO)

Algorithm	**Accuracy**			**F1 - Score**		
	AM	**SMOTE**	**Cast Iron**	**AM**	**SMOTE**	**Cast Iron**
Random Forest	*95%*	*93%*	*89%*	*46%*	*59%*	*56%*
XGBOOST	*95%*	*94%*	*91%*	*55%*	*61%*	*67%*
Logistic Regression	*94%*	*83%*	*83%*	*36%*	*42%*	*58%*
ANN	*95%*	*95%*	*89%*	*51%*	*52%*	*58%*

**** AM = All Materials,***

**** SMOTE = All metrials with SMOTE method***

Region of Waterloo

Region of Waterloo is another network that has been analyzed carefully. This network includes 4,517 pipes with different characteristics such as material, length, diameter, lining status, lining material, lining age, age, and the target variable.

Among non-broken pipes, PVC is the most frequent type with almost 50% of the total, followed by ductile iron, concrete, cast iron, and asbestos cement with 27.83%, 8.40%, 7.29%, and 5.39%, respectively. For broken pipes, however, ductile iron experienced more failures with a 37.76% contribution. Cast iron also accounts for almost 32.65% of broken pipes. It should be noted that the number of broken pipes compared to non-broken pipes is significantly fewer, and this shape of the dataset caused significant challenges while learning the models. The given figure indicates the frequency of each material based on different break statuses (Figure 0.8).

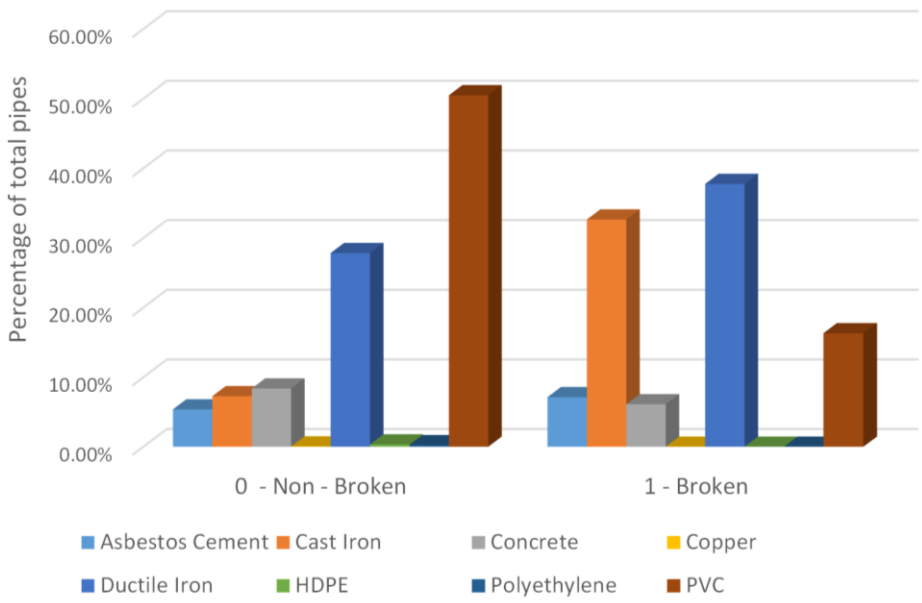


FIGURE 0.8 - PERCENTAGE OF EACH MATERIAL FOR BOTH CLASSES (0,1) BASED ON TOTAL PERCENTAGE OF EACH CLASS (REGION OF WATERLOO)

As previously discussed in previous chapters, having an imbalanced dataset may cause some challenges for various algorithms during the learning process. For example, there is a minority class and a majority class in most classification problems. Should the rate of these classes be worse than 1:10, then that dataset can be considered imbalanced.

Here, the Region of Waterloo has such a structure based on the information, with a rate of 1:50. From all pipes, 4,419 are reported as non-broken, and merely 98 unique pipes were reported as broken. Turning this dataset into a train, validation, and test set would result in 23 broken pipes and 881 non-broken pipes, leading to difficulty for the algorithms to learn a logical and satisfactory pattern from broken pipes.

The given confusion matrix compares the results of all four models for the Region of Waterloo network. It can be seen that random forest, XGBOOST, and ANN could not correctly predict the broken pipes, and there is a significant misclassification rate. The logistic regression model, however, was able to predict 16 broken pipes out of 23. This indicates that in some cases, where the dataset is significantly imbalanced, this algorithm is more powerful to detect broken pipes. Nonetheless, it was not able to detect the non-broken pipes correctly. Overall, ANN indicated a better performance in the prediction of both classes with an F1-Score of 19%. However, when SMOTE method was applied to overcome challenges related to imbalanced format, XGBOOST indicated a better performance with a 28% F1-Score. Finally, random forest with an F1-Score of 57% for cast iron pipes and Accuracy of 92% was the best predictive model. Other values can be found in the given tables and the appendix (TABLE 0.11; TABLE 0.12).

TABLE 0.110.11 - CONFUSION MATRIX FOR ALL MATERIALS (REGION OF WATERLOO)

Random Forest		Predicted		XGBOOST		Predicted	
		0	1			0	1
Actual	1	*22*	*1*	Actual	1	*22*	*1*
	0	*881*	*0*		0	*877*	*4*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*7*	*16*		**1**	*20*	*3*
	0	*628*	*253*	**Actual**	**0**	*876*	*5*

0 = None-Broken
1 = Broken

TABLE 0.120.12 - CLASSIFICATION RESULTS (REGION OF WATERLOO)

Algorithm	Accuracy			F1 - Score		
	AM	SMOTE	Cast Iron	AM	SMOTE	Cast Iron
Random Forest	*98%*	*92%*	*92%*	*8%*	*21%*	*57%*
XGBOOST	*96%*	*96%*	*87%*	*7%*	*28%*	*31%*
Logistic Regression	*71%*	*72%*	*72%*	*11%*	*11%*	*33%*
ANN	*97%*	*97%*	*86%*	*19%*	*19%*	*38%*

**** AM = All Materials,***
**** SMOTE = All metrials with SMOTE method***

Region of Durham

Region of Durham, with 21,344 pipes within the prepared classification file, is among the most extensive networks. In addition, there are different input variables such as material, lining material, lining status, protection status, length, diameter, age, lining age, protection age, and dependent variable, the target.

Same as the majority of the networks, PVC is the most frequent type among non-broken pipes, with almost 70% of the whole pipes. However, cast iron is the pipe the experienced failure more frequently than other pipes, with a 54.87% contribution among broken pipes. Ductile iron is another material that makes up almost 35% of the total recorded failures. The contribution of other materials is shown in the given figure (Figure 0.9).

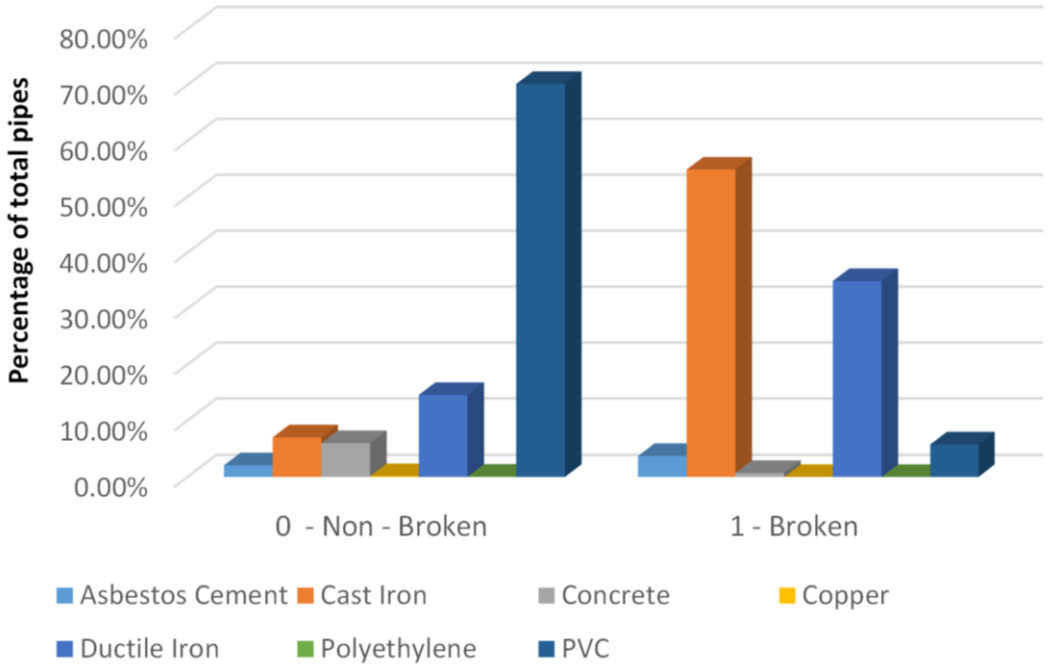


FIGURE 0.9 - PERCENTAGE OF EACH MATERIAL FOR BOTH CLASSES (0,1) BASED ON TOTAL PERCENTAGE OF EACH CLASS (REGION OF DURHAM)

From the entire available pipes, 19,332 are reported as non-broken and 2,012 as broken. Therefore, from this amount, 3,887 and 382 pipes are non-broken and broken, respectively, and they belong to the 20% test set. Evaluation of the model on the test indicates satisfactory results for all categories. With 297 correctly classified broken pipes, XGBOOST showed the best performance among these classifiers, although the results are relatively similar. For instance, random forest and XGBOOST with an F1-Score of 85% are the best classifiers. Moreover, the accuracy of these two models is 98% and 97%, respectively. On the other hand, random forest and XGBOOST have the highest F1-Score for cast iron pipes which is 89%, and ANN, with only 1% difference, with the F1-Score of 88%, is in the following position (TABLE 0.13, TABLE 0.14).

TABLE 0.130.13 - CONFUSION MATRIX FOR ALL MATERIALS (REGION OF DURHAM)

Random Forest	Predicted		XGBOOST	Predicted	
	0	1		0	1

Actual	**1**	*95*	*287*	**Actual**	**1**	*85*	*297*
	0	*3877*	*10*		**0**	*3865*	*22*

Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*110*	*272*		**1**	*88*	*294*
	0	*3866*	*21*	**Actual**	**0**	*3858*	*29*

0 = None-Broken
1 = Broken

TABLE 0.140.14 - CLASSIFICATION RESULTS (REGION OF DURHAM)

Algorithm	**Accuracy**			**F1 - Score**		
	AM	**SMOTE**	**Cast Iron**	**AM**	**SMOTE**	**Cast Iron**
Random Forest	*98%*	*96%*	*91%*	*85%*	*80%*	*89%*
XGBOOST	*97%*	*97%*	*91%*	*85%*	*82%*	*89%*
Logistic Regression	*97%*	*91%*	*89%*	*81%*	*65%*	*86%*
ANN	*97%*	*97%*	*90%*	*83%*	*83%*	*88%*

**** AM = All Materials,***
**** SMOTE = All metrials with SMOTE method***

Calgary

After making the classification dataset for Calgary, the number of pipes is 55,462, including broken and non-broken pipes. PVC is the most frequent pipe for non-broken pipes with 60.88% of class 0. DI and CI follow this material with 18.89% and 10.45% contributions, respectively. For class 1 or broken pipes, however, cast iron with 61.36% experienced the most failures. In addition, 32.75% of the broken pipes are related to ductile iron pipes (Figure 0.10).

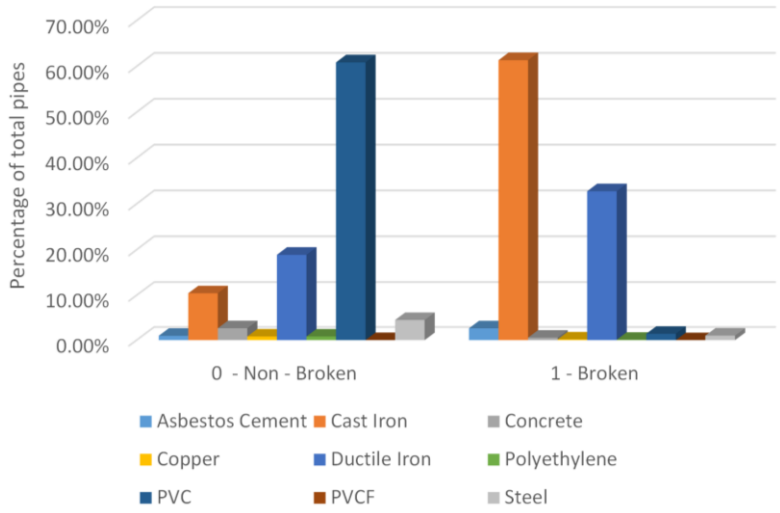


FIGURE 0.10 - PERCENTAGE OF EACH MATERIAL FOR BOTH CLASSES (0,1) BASED ON TOTAL PERCENTAGE OF EACH CLASS (CALGARY)

Based on the available information, 51,029 pipes are non-broken, and 4,433 are reported to have broken. Therefore, 20% was selected for evaluation as a test size, including 10,211 non-broken and 882 broken pipes. Once more, XGBOOST indicated a better performance based on the given confusion matrix, with an accuracy score of 98% and an F1-Score of 88%, for all materials (TABLE 0.15). However, when the SMOTE method was used, ANN represented better performance with an F1-Score of 89%, as opposed to other models that experienced a decline in the overall performance. Finally, random forest, XGBOOST, and ANN performed similarly with an F1-Score of 90% for cast iron pipes. It should be noted that cast iron pipes results related to logistic regression were also desirable with an F1-Score of 86%.

TABLE 0.150.15 - CONFUSION MATRIX FOR ALL MATERIALS (CALGARY)

Random Forest		**Predicted**		**XGBOOST**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*189*	*693*	**Actual**	**1**	*157*	*725*
	0	*10181*	*30*		**0**	*10167*	*44*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*257*	*625*		**1**	*169*	*713*
	0	*10172*	*39*	**Actual**	**0**	*10171*	*40*

0 = None-Broken, 1 = Broken

TABLE 0.160.16 - CLASSIFICATION RESULTS (CALGARY)

Algorithm	Accuracy			F1 - Score		
	AM	SMOTE	Cast Iron	AM	SMOTE	Cast Iron
Random Forest	*98%*	*95%*	*93%*	*86%*	*74%*	*90%*
XGBOOST	*98%*	*96%*	*93%*	*88%*	*80%*	*90%*
Logistic Regression	*97%*	*90%*	*91%*	*81%*	*58%*	*86%*
ANN	*98%*	*98%*	*93%*	*87%*	*87%*	*90%*

**** AM = All Materials,***
**** SMOTE = All metrials with SMOTE method***

Vancouver

Vancouver is another network with a significant imbalanced dataset, and consequently, the results are not satisfactory. This network consists of 63,236 pipes, most of which are non-broken, and only a small proportion of these pipes are broken. Different input variables are prepared for this network, such as diameter, length, coating material, lining material, age, and the target variable. The predominant materials within this network are ductile iron and cast iron, in both class 0 and class 1. Ductile iron is the most frequent type, with 55% non-broken pipes, followed by cast iron with 43.74%. Furthermore, most pipes within break records are made from cast iron, with almost 90% of total broken pipes (Figure 0.11).

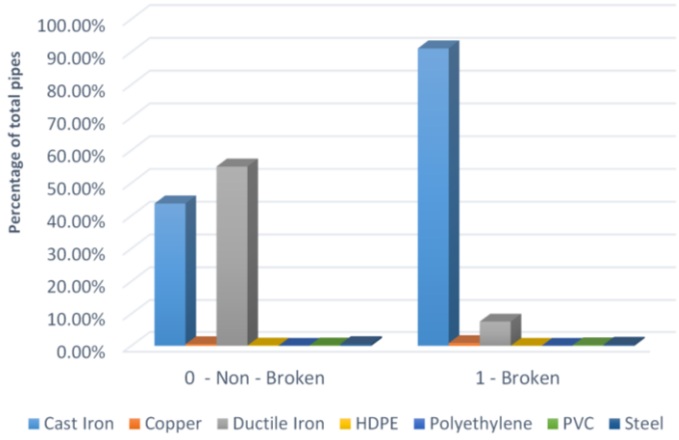


FIGURE 0.11 - PERCENTAGE OF EACH MATERIAL FOR BOTH CLASSES (0,1) BASED ON TOTAL PERCENTAGE OF EACH CLASS (VANCOUVER)

From the entire data points within this network, 62,654 pipes are non-broken, and only 582 are broken. Unfortunately, this makes Vancouver`s dataset significantly imbalanced, leading to a low-score performance. For the evaluation step, 125 broken pipes and 12,523 non-broken pipes were selected randomly. Learning a logical pattern for broken pipes is significantly complex for any algorithm in this case, where there is no sufficient information for one class.

As can be seen from the given table, random forest, XGBOOST, and ANN were able to predict only 20 broken pipes correctly while mispredicting 105 pipes incorrectly (TABLE 0.17). In this case, logistic regression, however, was able to predict 114 broken pipes out of 125 correctly while losing accuracy for the prediction of non-broken pipes, compared to other models. Eventually, XGBOOST and ANN, with 26% F1-Score, indicated a better performance than others (TABLE 0.18). Moreover, for cast iron also, XGBOOST was able to achieve 28% performance for the F1-score, which is not a desirable result for future prediction.

TABLE 0.170.17 - CONFUSION MATRIX FOR ALL MATERIALS (VANCOUVER)

Random Forest		**Predicted**		**XGBOOST**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*105*	*20*	**Actual**	**1**	*105*	*20*
	0	*12503*	*20*		**0**	*12513*	*10*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*11*	*114*		**1**	*105*	*20*
	0	*8362*	*4161*	**Actual**	**0**	*12517*	*6*

0 = None-Broken
1 = Broken

TABLE 0.180.18 - CLASSIFICATION RESULTS (VANCOUVER)

Algorithm	**Accuracy**			**F1 - Score**		
	AM	**SMOTE**	**Cast Iron**	**AM**	**SMOTE**	**Cast Iron**
Random Forest	*99%*	*82%*	*99%*	*16%*	*8%*	*26%*
XGBOOST	*99%*	*91%*	*99%*	*26%*	*12%*	*28%*
Logistic Regression	*67%*	*67%*	*65%*	*5%*	*5%*	*7%*
ANN	*99%*	*99%*	*98%*	*26%*	*26%*	*26%*

**** AM = All Materials,***

**** SMOTE = All metrials with SMOTE method***

Victoria

The final classification file of Victoria included 3,149 pipes. In addition, this network provided various variables such as material, diameter, HGL (Hydraulic Grade Line), length, lining material, lining status, age, and target. As previously mentioned in the literature, HGL and water pressure are among the attributes that should be studied in more detail since not much information has been collected regarding these attributes in most utilities in Canada. Nonetheless, this attribute was provided by Victoria.

As shown in the given graph, for the non-broken group (class 0), cast iron and ductile iron are the most frequent pipes for this utility, with almost 41% of the total for each (Figure 0.12). The other material in this category is PVC and HDPE, accounting for 8.94% and 3.47%, respectively. On the other hand, cast iron with 73.41% and ductile iron with 18.54% are the most frequent materials for broken pipes. The percentage of other materials can be found in the given bar graph (Figure 0.12).

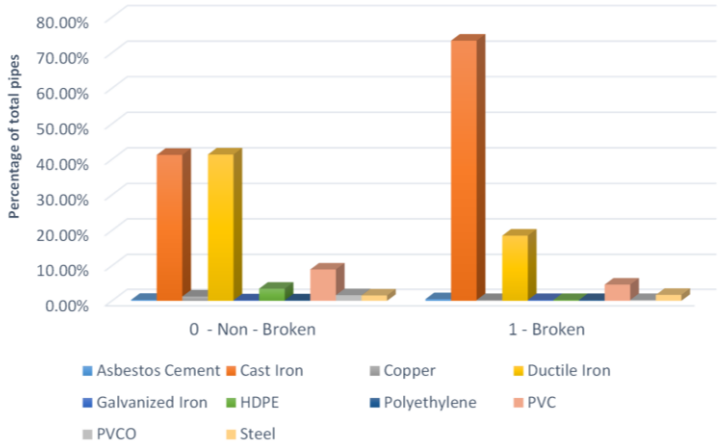


FIGURE 0.12 - PERCENTAGE OF EACH MATERIAL FOR BOTH CLASSES (0,1) BASED ON TOTAL PERCENTAGE OF EACH CLASS (VICTORIA)

In this network, there are 2,739 non-broken pipes and 410 broken pipes. Like other utilities, considering 20% for the test set, the number of broken pipes is 93, and non-broken 537.

Results for this utility did not show a good performance. However, XGBOOST, another time, was able to detect more broken pipes than other models and became the best classifier for this network. More details about the performance of these models are found in the given confusion matrix (TABLE 0.19).

TABLE 0.190.19 - CONFUSION MATRIX FOR ALL MATERIALS (VICTORIA)

Random Forest		Predicted		XGBOOST		Predicted	
		0	1			0	1
Actual	1	*60*	*33*	Actual	1	*56*	*37*
	0	*535*	*2*		0	*532*	*5*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*63*	*30*		**1**	*59*	*34*
	0	*526*	*11*	**Actual**	**0**	*524*	*13*

0 = None-Broken
1 = Broken

XGBOOST, with an accuracy of 90% and F1-Score of 55%, has proven to be the best algorithm for all materials. For the SMOTE method and the cast iron group, this algorithm demonstrated the highest performance compared to others, with an F1-Score of 57% and 70%, in successive. The following table shows the accuracy and F1-Score for all algorithms (TABLE 0.20).

TABLE 0.200.20 - CLASSIFICATION RESULTS (VICTORIA)

Algorithm	Accuracy			F1 - Score		
	AM	SMOTE	Cast Iron	AM	SMOTE	Cast Iron
Random Forest	*90%*	*82%*	*88%*	*52%*	*53%*	*69%*
XGBOOST	*90%*	*89%*	*87%*	*55%*	*57%*	*70%*
Logistic Regression	*88%*	*79%*	*85%*	*45%*	*53%*	*59%*
ANN	*89%*	*89%*	*87%*	*49%*	*49%*	*68%*

**** AM = All Materials,***
**** SMOTE = All metrials with SMOTE method***

Halifax

The final classification file for Halifax, after cleaning and merging, consists of 12,999 pipes. This file has different variables such as material, length diameter, lining status, lining material, age, and the target variable. Given bar chart compares different materials in both class 0 and class 1 (Figure 0.13). Ductile iron and cast iron are the most frequent materials in both classes. For non-broken pipes, ductile iron accounts for almost 67% of the total, and cast iron makes up 21%. However, cast iron has the most recorded failures for broken pipes, with over 83% of the total. This material is followed by cast iron with just over 10% of total failures.

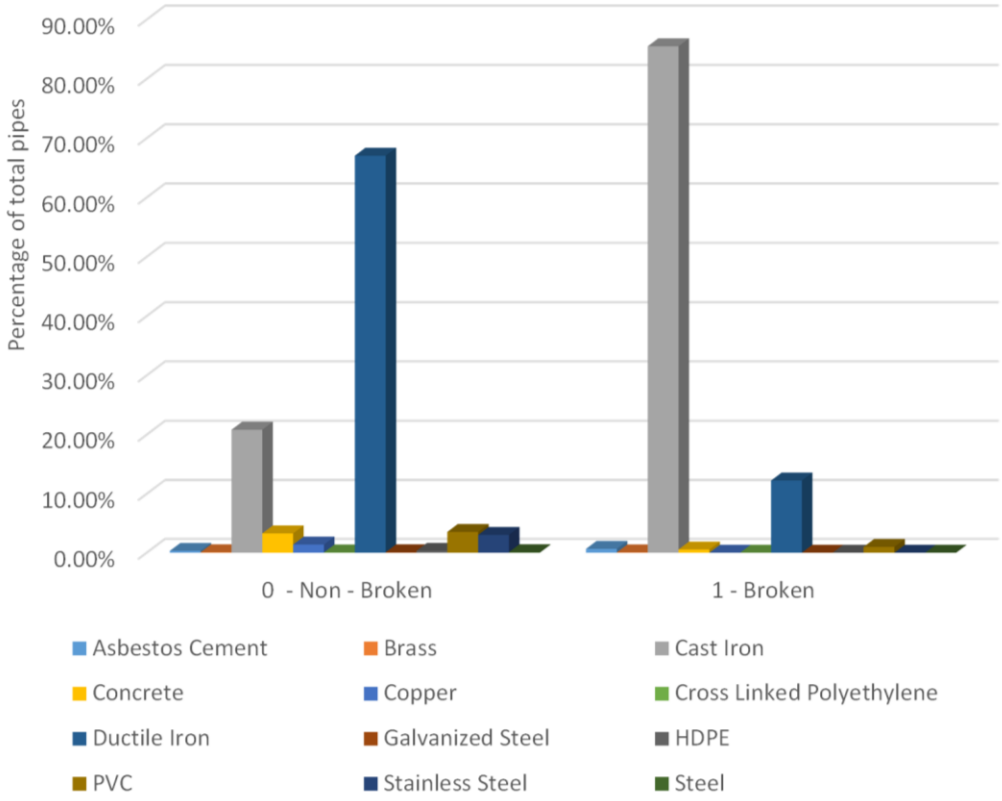


FIGURE 0.13 - PERCENTAGE OF EACH MATERIAL FOR BOTH CLASSES (0,1) BASED ON TOTAL PERCENTAGE OF EACH CLASS (HALIFAX)

From all pipes within the network, 11,164 pipes are in class 0 and 1,835 in class 1. Moreover, based on the 20% assumption, there are 369 broken and 2,231 non-broken pipes in the test set. This time, almost all algorithms indicated satisfactory performance. Random forest, XGBOOST, and ANN were able to predict 264 broken pipes correctly, which resulted in 95% accuracy and 79% F1-Score. Logistic regression with 75% F1-Score was the weakest model (TABLE 0.21; TABLE 0.22). When SMOTE method was applied to all materials, the accuracy of all models worsened, but not ANN, which remained at 79%. For cast-iron pipes, however, the accuracy of all models decreased while the F1-Score improved. ANN was the best algorithm for the group of cast iron pipes with an 86% F1-Score.

TABLE 0.210.21 - CONFUSION MATRIX FOR ALL MATERIALS (HALIFAX)

Random Forest		Predicted		XGBOOST		Predicted	
		0	1			0	1
Actual	1	*115*	*254*	Actual	1	*105*	*264*
	0	*2207*	*24*		0	*2194*	*37*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*118*	*251*		**1**	*105*	*264*
	0	*2184*	*47*	**Actual**	**0**	*2198*	*33*

0 = None-Broken
1 = Broken

TABLE 0.220.22 - CLASSIFICATION RESULTS (HALIFAX)

Algorithm	Accuracy			F1 - Score		
	AM	SMOTE	Cast Iron	AM	SMOTE	Cast Iron
Random Forest	*95%*	*93%*	*89%*	*79%*	*76%*	*85%*
XGBOOST	*95%*	*93%*	*87%*	*79%*	*78%*	*82%*
Logistic Regression	*94%*	*85%*	*89%*	*75%*	*61%*	*85%*
ANN	*95%*	*95%*	*90%*	*79%*	*79%*	*86%*

**** AM = All Materials,***
**** SMOTE = All metrials with SMOTE method***

St. John's

St. John's is another utility in this study, including 8,863 pipes after cleaning and preparing the classification dataset. In addition, this city provided a variety of attributes such as material, diameter, roughness, length, age, and the target variable.

Three materials are the predominant types of the entire network for non-broken pipes. First, ductile iron with 48.95% is the most frequent material within the network for class 0. This material is then followed by cast iron with 36.15% and PVC pipe with 13.99% contribution to

the total non-broken pipes. Broken pipes, however, follow an inverse pattern, with having cast iron as the most frequent material that has experienced more failure in the network, with 83% of failure records. Finally, ductile iron is another frequent material that accounts for almost 15.25% of total failures (Figure 0.14).

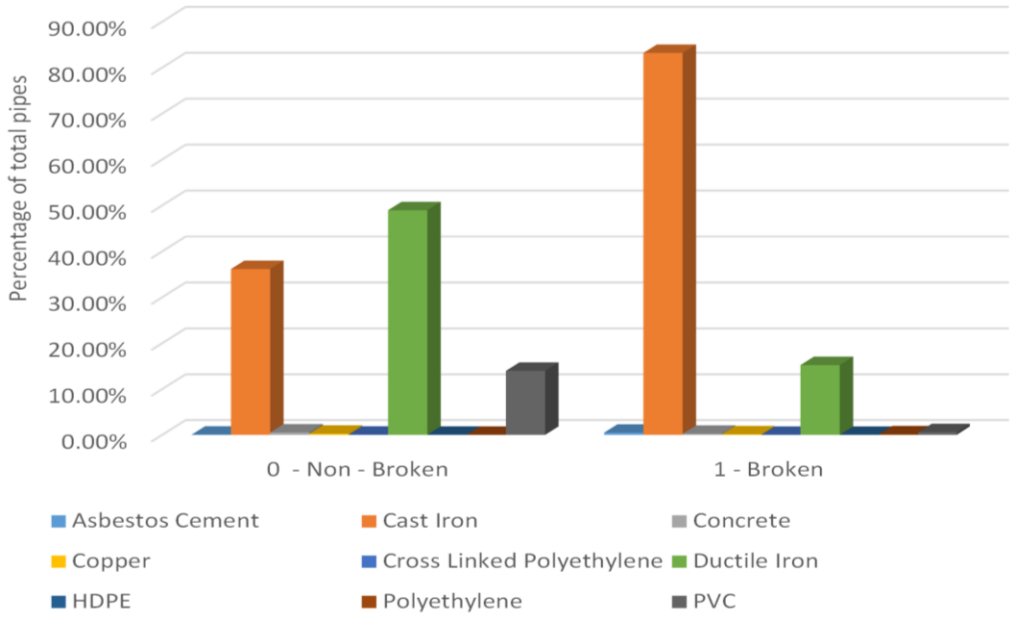


FIGURE 0.14 - PERCENTAGE OF EACH MATERIAL FOR BOTH CLASSES (0,1) BASED ON TOTAL PERCENTAGE OF EACH CLASS (ST. JOHN'S)

For class 0 and class 1 of the classification file, there are different numbers of pipes. For non-broken, 8,030 pipes, and for broken pipes, 833 pipes are recorded in the file. After creating a test split for the evaluation process, 170 broken and 1,603 non-broken pipes can be found. The Given confusion matrix was prepared based on the test set. XGBOOST, similar to most cities that have been explained thus far, detected the most broken pipes, with 89 correctly classified in class 1.

Interestingly, comparing the results from all cities indicates that random forest seems to be able to detect the most number of non-broken pipes among all classifiers. For instance, in this case, this algorithm was able to classify 1,594 non-broken pipes correctly and merely nine pipes incorrectly. However, overall, all models did not show relatively satisfactory results, based on the classification metrics.

TABLE 0.230.23 - CONFUSION MATRIX FOR ALL MATERIALS (ST. JOHNS'S)

Random Forest		**Predicted**		**XGBOOST**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*107*	*63*	**Actual**	**1**	*81*	*89*
	0	*1594*	*9*		**0**	*1584*	*19*

Logistic Regression		Predicted		ANN		Predicted	
		0	1			0	1
Actual	1	*47*	*123*		1	*100*	*70*
	0	*1257*	*346*	Actual	0	*1576*	*27*

0 = None-Broken, 1 = Broken

XGBOOST for the first category (all materials), SMOTE, and cast iron have proven to be the best algorithm with 94%, 91%, and 90%, respectively (TABLE 0.24). However, as the dataset is imbalanced, F1-Score is of significant importance. This metric for XGBOOST, is 64% for the first category, 58% for SMOTE, and 72% for cast iron pipes. The critical point here is that making a homogenous group of pipes would lead to higher accuracy. Some of the other materials in the network were also analyzed separately. Related results are provided in the appendix, and it makes the comparison easier. The given table indicates the accuracy and F1-Score for all models in the claasification step (TABLE 0.24).

TABLE 0.240.24 - CLASSIFICATION RESULTS (ST. JOHNS'S)

Algorithm	Accuracy			F1 - Score		
	AM	SMOTE	Cast Iron	AM	SMOTE	Cast Iron
Random Forest	*93%*	*84%*	*89%*	*52%*	*47%*	*63%*
XGBOOST	*94%*	*91%*	*90%*	*64%*	*58%*	*72%*
Logistic Regression	*78%*	*78%*	*68%*	*38%*	*39%*	*46%*
ANN	*93%*	*93%*	*88%*	*52%*	*54%*	*59%*

**** AM = All Materials,***
**** SMOTE = All metrials with SMOTE method***

Barrie

This network consists of 5,183 pipes for the classification file. Material, diameter, protection status, length, casing material, restrained, age, and the target are the variables prepared for this dataset.

For class 0 of this study, PVC is the most frequent type of material. This material contributes to almost 68% of pipes that experienced no failure, and ductile iron follows PVC with around 23.43%. Cast iron and copper pipes are the other utilized materials equally distributed among non-broken pipes, with almost 4% contribution.

For broken pipes, on the other hand, the same trend as some of the other utilities can be noticed. Cast iron, once more, is the material that has experienced at least one failure more than any other materials within Barrie`s network. This material accounts for just over 60% of class 1. Furthermore, ductile iron is the second material that makes up 26.25% of total broken pipes. PVC and copper are the other materials with a relatively small proportion of the total broken pipes (Figure 0.15).

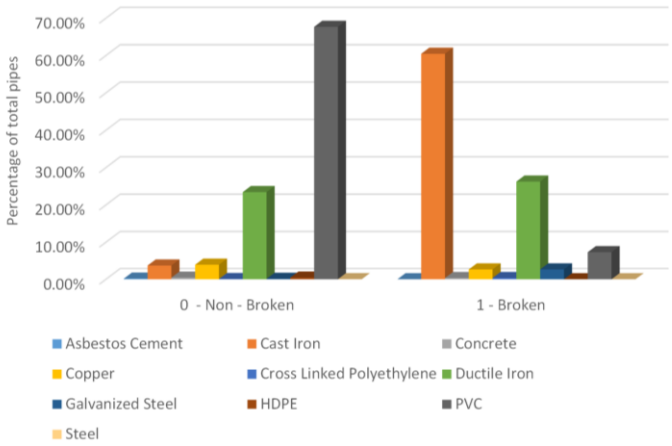


FIGURE 0.15 - PERCENTAGE OF EACH MATERIAL FOR BOTH CLASSES (0,1) BASED ON TOTAL PERCENTAGE OF EACH CLASS (BARRIE)

Among the entire pipes, 4,882 are non-broken, and 301 pipes experienced at least one failure. From this number, 61 broken pipes and 976 non-broken pipes belong to the test set for the evaluation process. From the prepared confusion matrix, ANN and XGBOOST were able to predict the same number of broken pipes correctly. However, ANN was able to find more non-broken pipes based upon the analysis. As a result, the accuracy for XGBOOST and ANN is 97%, and the F1-Scores are 71% and 73%, respectively, indicating a better performance for the ANN algorithm. ANN was also the best predictive model for SMOTE method with an F1-Score of 73%. However, for cast-iron pipes, random forest and logistic regression represented a better performance with an accuracy of 85% and F1-Score of 85% for both models. Given tables provides more information regarding the results, and more results can be found in the appendix (TABLE 0.25; TABLE 0.26).

TABLE 0.250.25 - CONFUSION MATRIX FOR ALL MATERIALS (BARRIE)

Random Forest		**Predicted**		**XGBOOST**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*23*	*38*	**Actual**	**1**	*21*	*40*
	0	*969*	*7*		**0**	*964*	*12*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	

		0	1			0	1
Actual	**1**	*26*	*35*		**1**	*21*	*40*
	0	*969*	*7*	**Actual**	**0**	*967*	*9*

0 = None-Broken
1 = Broken

TABLE 0.260~~0.26~~ - CLASSIFICATION RESULTS (BARRIE)

Algorithm	**Accuracy**			**F1 - Score**		
	AM	**SMOTE**	**Cast Iron**	**AM**	**SMOTE**	**Cast Iron**
Random Forest	*97%*	*96%*	*85%*	*72%*	*70%*	*85%*
XGBOOST	*97%*	*95%*	*82%*	*71%*	*67%*	*83%*
Logistic Regression	*97%*	*89%*	*85%*	*68%*	*49%*	*85%*
ANN	*97%*	*97%*	*78%*	*73%*	*73%*	*79%*

**** AM = All Materials,***
**** SMOTE = All metrials with SMOTE method***

Feature Importance Figures

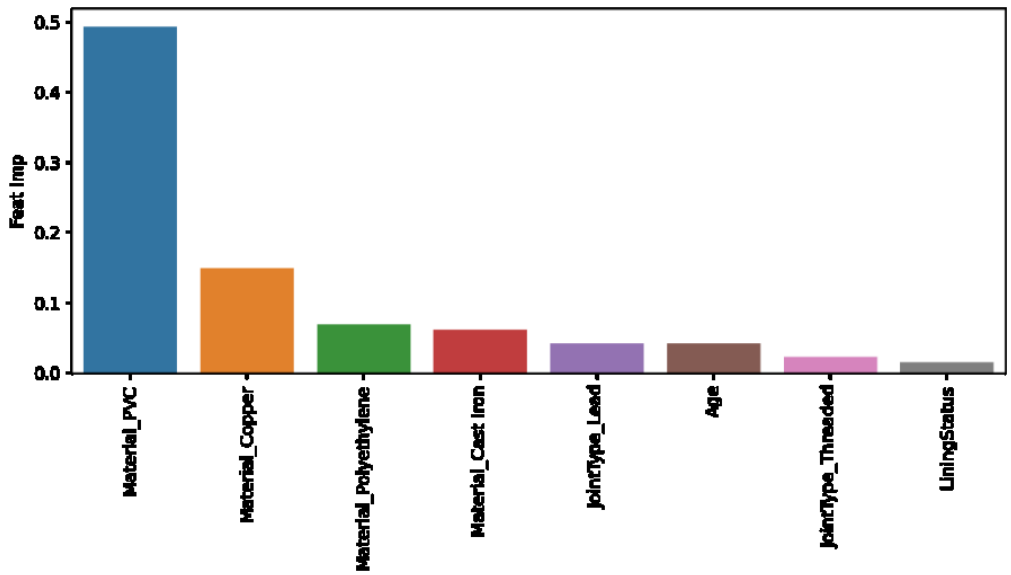


FIGURE 0.16 - THE MOST IMPORTANT FEATURES BASED ON XGBOOST RESULTS (SASKATOON)

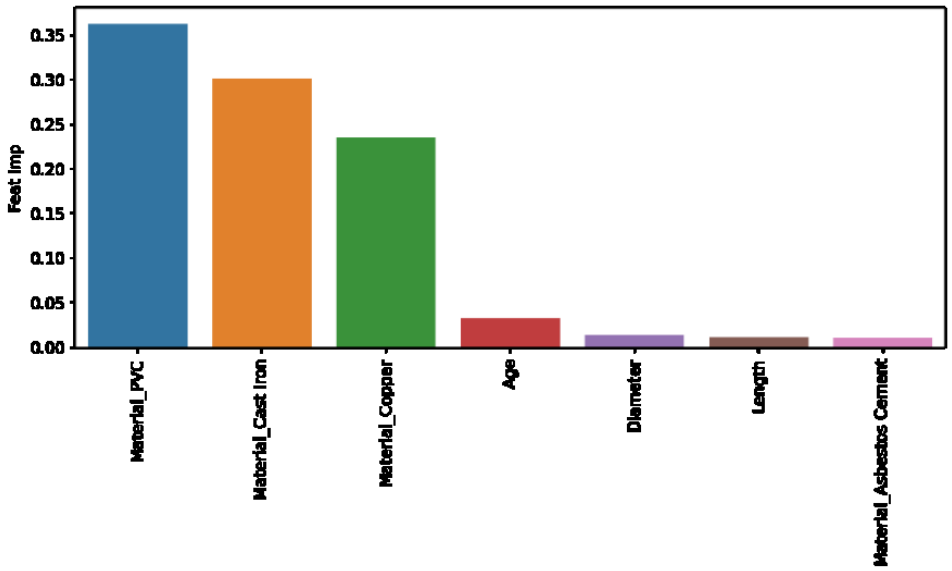


FIGURE 0.17 - THE MOST IMPORTANT FEATURES BASED ON XGBOOST RESULTS (WINNIPEG)

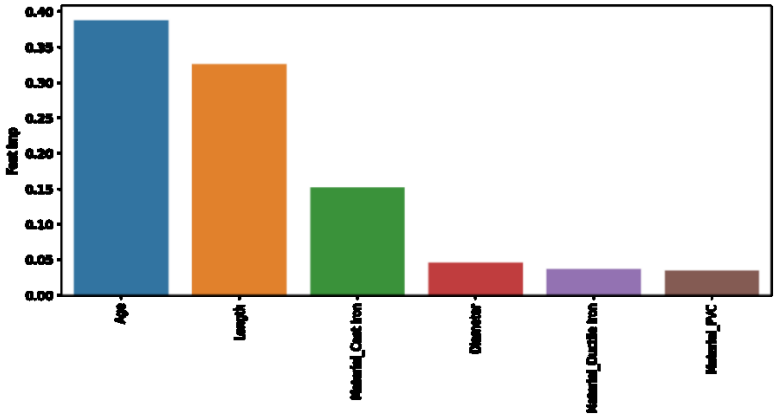


FIGURE 0.18 – THE MOST IMPORTANT FEATURES BASED ON XGBOOST RESULTS (KITCHENER)

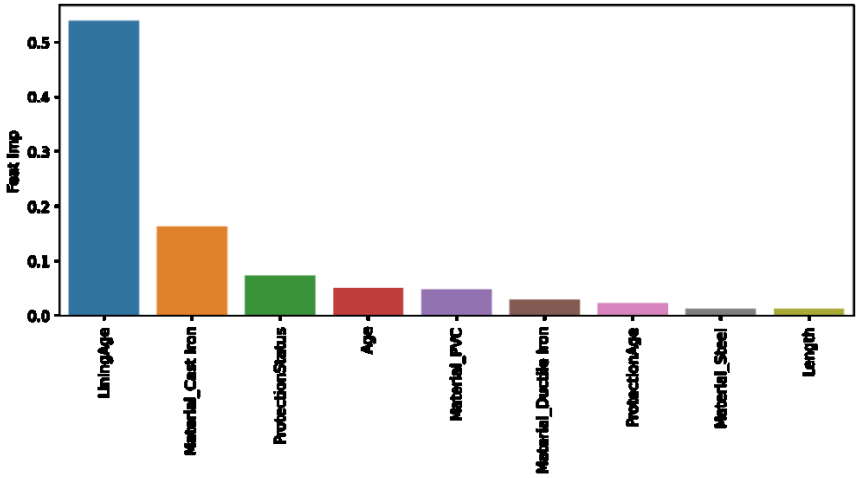


FIGURE 0.19 - THE MOST IMPORTANT FEATURES BASED ON XGBOOST RESULTS (MARKHAM)

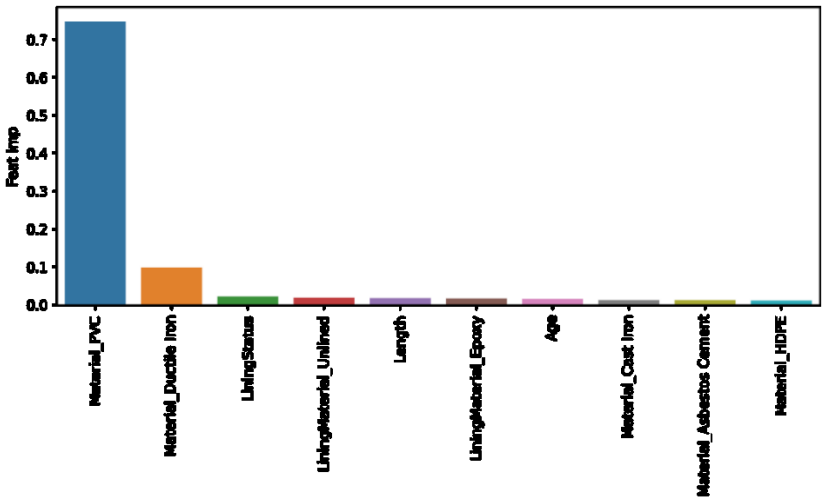


FIGURE 0.20 - THE MOST IMPORTANT FEATURES BASED ON XGBOOST RESULTS (WATERLOO)

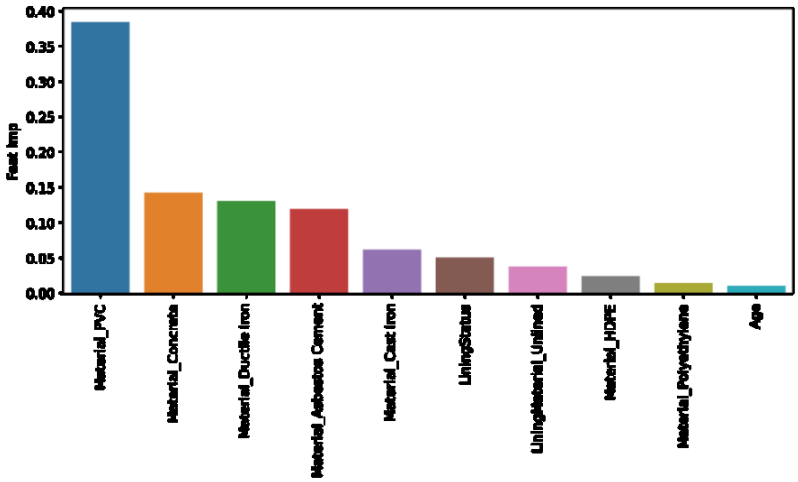


FIGURE 0.21 - THE MOST IMPORTANT FEATURES BASED ON XGBOOST RESULTS (REGION OF WATERLOO)

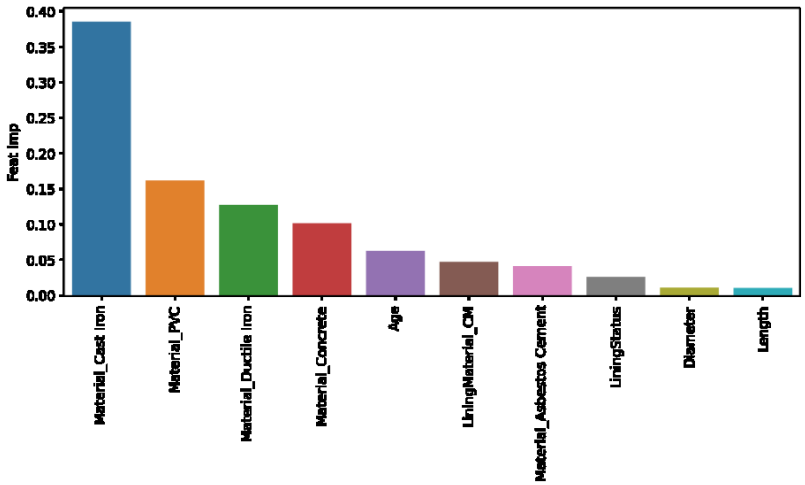


FIGURE 0.22 - THE MOST IMPORTANT FEATURES BASED ON XGBOOST RESULTS (REGION OF DURHAM)

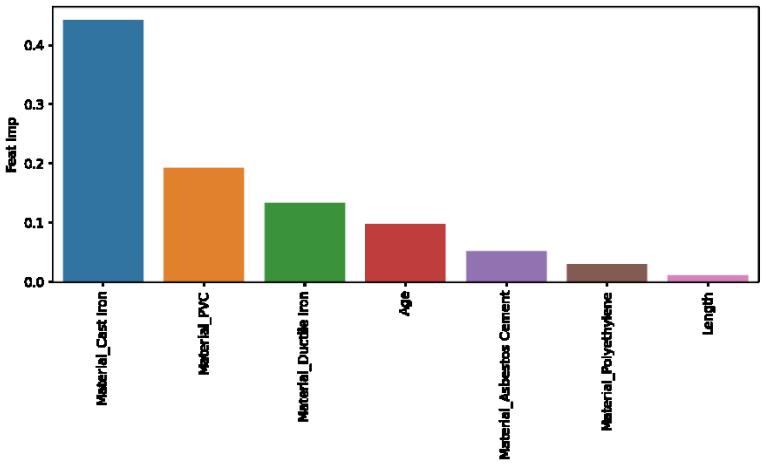


FIGURE 0.23 - THE MOST IMPORTANT FEATURES BASED ON XGBOOST RESULTS (CALGARY)

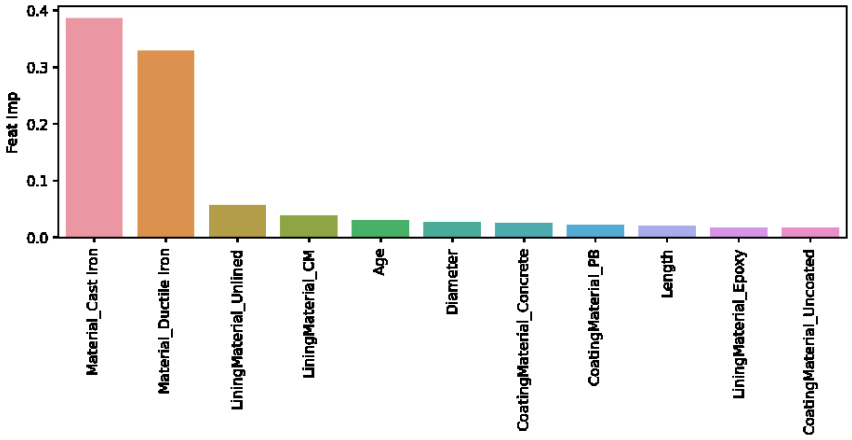


FIGURE 0.24 - THE MOST IMPORTANT FEATURES BASED ON XGBOOST RESULTS (VANCOUVER)

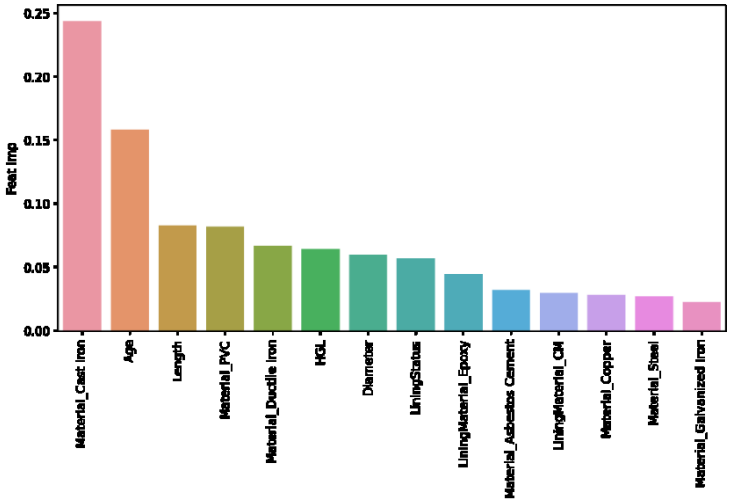


FIGURE 0.25 - THE MOST IMPORTANT FEATURES BASED ON XGBOOST RESULTS (VICTORIA)

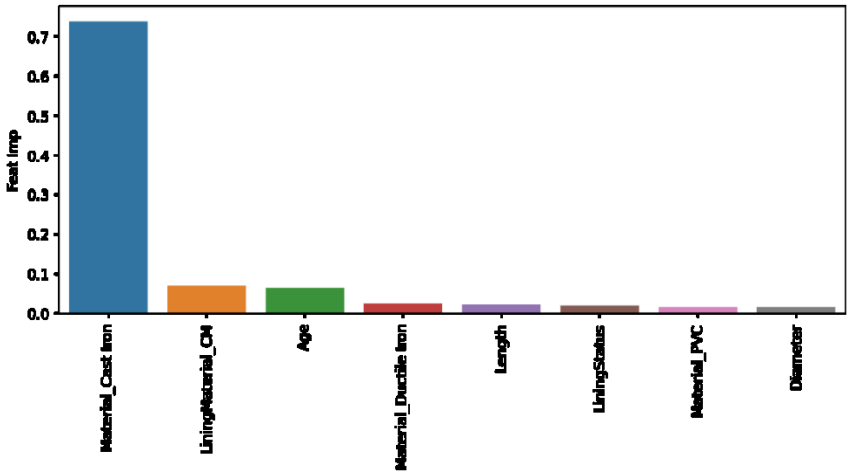


FIGURE 0.26 - THE MOST IMPORTANT FEATURES BASED ON XGBOOST RESULTS (HALIFAX)

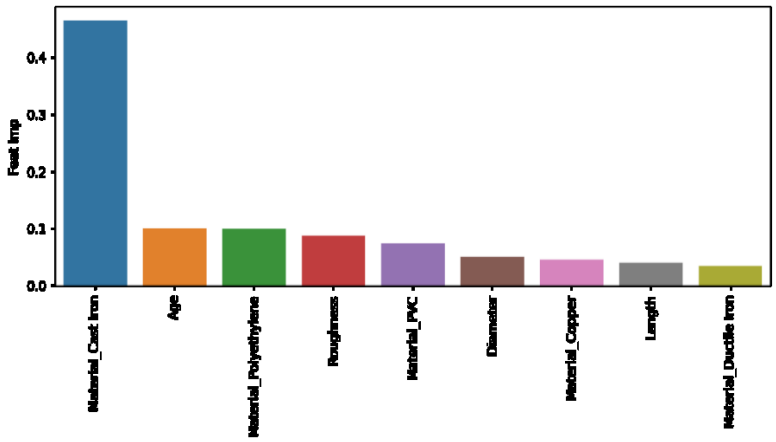


FIGURE 0.27 - THE MOST IMPORTANT FEATURES BASED ON XGBOOST RESULTS (ST. JOHNS'S)

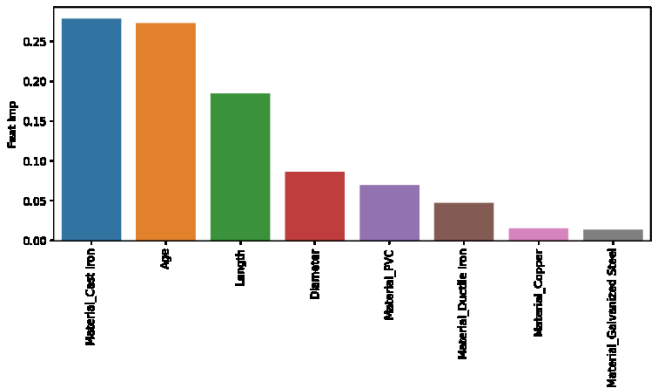


FIGURE 0.28 - THE MOST IMPORTANT FEATURES BASED ON XGBOOST RESULTS (BARRIE)

APPENDIX D - CONFUSION MATRIX FOR SMOTE METHOD AND CAST IRON PIPES

TABLE 0.1~~0.1~~ - CONFUSION MATRIX FOR ALL MATERIALS – SMOTE METHOD (SASKATOON)

Random Forest		**Predicted**		**XGBOOST**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*105*	626	**Actual**	**1**	*114*	*617*
	0	*5529*	202		**0**	*5597*	*134*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**

Actual	**1**	*84*	*647*		**1**	*153*	*578*
	0	*5144*	*587*	**Actual**	**0**	*5675*	*56*

0 = None-Broken
1 = Broken

TABLE 0.2~~0.2~~ - CONFUSION MATRIX FOR CAST IRON (SASKATOON)

Random Forest		**Predicted**		**XGBOOST**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*67*	*294*	**Actual**	**1**	*52*	*309*
	0	*797*	*18*		**0**	*793*	*22*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*65*	*296*		**1**	*62*	*299*
	0	*712*	*103*	**Actual**	**0**	*783*	*32*

0 = None-Broken
1 = Broken

TABLE 0.3~~0.3~~ - CONFUSION MATRIX FOR ALL MATERIALS – SMOTE METHOD (WINNIPEG)

Random Forest		**Predicted**		**XGBOOST**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*205*	1424	**Actual**	**1**	*260*	*1369*
	0	*17366*	1532		**0**	*17852*	*1046*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*176*	*1453*		**1**	*577*	*1052*
	0	*16370*	*2528*	**Actual**	**0**	*18709*	*189*

0 = None-Broken
1 = Broken

TABLE 0.4~~0.4~~ - CONFUSION MATRIX FOR CAST IRON (WINNIPEG)

Random Forest		Predicted		XGBOOST		Predicted	
		0	**1**			**0**	**1**
Actual	**1**	*394*	*704*	**Actual**	**1**	*351*	*747*
	0	*2580*	*110*		**0**	*2554*	*136*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*298*	*800*		**1**	*366*	*732*
	0	*2134*	*556*	**Actual**	**0**	*2531*	*159*

0 = None-Broken
1 = Broken

TABLE 0.5~~0.5~~ - CONFUSION MATRIX FOR ALL MATERIALS – SMOTE METHOD (KITCHENER)

Random Forest		Predicted		XGBOOST		Predicted	
		0	**1**			**0**	**1**
Actual	**1**	*30*	*180*	**Actual**	**1**	*43*	*167*
	0	*2319*	*385*		**0**	*2463*	*241*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*32*	*178*		**1**	*100*	*110*
	0	*2240*	*464*	**Actual**	**0**	*2689*	*15*

0 = None-Broken
1 = Broken

TABLE 0.60.6 - CONFUSION MATRIX FOR CAST IRON (KITCHENER)

Random Forest		Predicted		XGBOOST		Predicted	
		0	1			0	1
Actual	1	*38*	*94*	Actual	1	*35*	*97*
	0	*440*	*17*		0	*438*	*19*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	1	*40*	*92*		1	*34*	*98*
	0	*430*	*27*	Actual	0	*434*	*23*

0 = None-Broken
1 = Broken

TABLE 0.70.7 - CONFUSION MATRIX FOR ALL MATERIALS – SMOTE METHOD (MARKHAM)

Random Forest		Predicted		XGBOOST		Predicted	
		0	1			0	1
Actual	1	*25*	103	Actual	1	27	101
	0	*2012*	18		0	2006	24
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	1	18	110		1	26	102
	0	1953	77	Actual	0	2023	7

0 = None-Broken
1 = Broken

TABLE 0.80.8 - CONFUSION MATRIX FOR CAST IRON (MARKHAM)

Random Forest		Predicted		XGBOOST		Predicted	
		0	1			0	1
Actual	1	*4*	*38*	Actual	1	*4*	*38*

	0	*26*	*0*		**0**	*26*	*0*

Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*3*	*39*		**1**	*3*	*39*
	0	*26*	*0*	**Actual**	**0**	*24*	*2*

0 = None-Broken
1 = Broken

TABLE 0.9~~0.9~~ - CONFUSION MATRIX FOR ALL MATERIALS – SMOTE METHOD (WATERLOO)

Random Forest		**Predicted**		**XGBOOST**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*27*	78	**Actual**	**1**	33	72
	0	*1320*	82		**0**	1342	60

Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	15	90		**1**	65	40
	0	1167	235	**Actual**	**0**	1393	9

0 = None-Broken
1 = Broken

TABLE 0.10~~0.10~~ - CONFUSION MATRIX FOR CAST IRON (WATERLOO)

Random Forest		**Predicted**		**XGBOOST**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*42*	*31*	**Actual**	**1**	*31*	*42*
	0	*362*	*7*		**0**	*359*	*10*

Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*20*	*53*		**1**	*41*	*32*

	0	*312*	*57*	**Actual**	**0**	*363*	*6*

0 = None-Broken
1 = Broken

TABLE 0.110~~0.11~~ - CONFUSION MATRIX FOR ALL MATERIALS – SMOTE METHOD (REGION OF WATERLOO)

Random Forest		**Predicted**		**XGBOOST**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*14*	9	**Actual**	**1**	16	7
	0	*827*	54		**0**	861	20
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	7	16		**1**	20	3
	0	632	249	**Actual**	**0**	876	5

0 = None-Broken
1 = Broken

TABLE 0.120~~0.12~~ - CONFUSION MATRIX FOR CAST IRON (REGION OF WATERLOO)

Random Forest		**Predicted**		**XGBOOST**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*5*	*4*	**Actual**	**1**	*7*	*2*
	0	*61*	*1*		**0**	*60*	*2*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*4*	*5*		**1**	*58*	*4*
	0	*46*	*16*	**Actual**	**0**	*6*	*3*

0 = None-Broken

1 = Broken

TABLE 0.130.13 - CONFUSION MATRIX FOR ALL MATERIALS – SMOTE METHOD (REGION OF DURHAM)

Random Forest		Predicted		XGBOOST		Predicted	
		0	1			0	1
Actual	1	*54*	328	Actual	1	*69*	*313*
	0	*3772*	115		0	*3821*	*66*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*44*	*338*		**1**	*88*	*294*
	0	*3567*	*320*	**Actual**	**0**	*3858*	*29*

0 = None-Broken
1 = Broken

TABLE 0.140.14 - CONFUSION MATRIX FOR CAST IRON (REGION OF DURHAM)

Random Forest		Predicted		XGBOOST		Predicted	
		0	1			0	1
Actual	1	*39*	*167*	Actual	1	*32*	*174*
	0	*283*	*4*		0	*276*	*11*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*34*	*172*		**1**	*32*	*174*
	0	*266*	*21*	**Actual**	**0**	*272*	*15*

0 = None-Broken
1 = Broken

TABLE 0.150.15 - CONFUSION MATRIX FOR ALL MATERIALS – SMOTE METHOD (CALGARY)

Random Forest		Predicted		XGBOOST		Predicted	
		0	1			0	1

Actual	**1**	*75*	*807*	**Actual**	**1**	*102*	*780*
	0	*9720*	*491*		**0**	*9917*	*294*

Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*78*	*804*		**1**	*167*	*715*
	0	*9130*	*1081*	**Actual**	**0**	*10163*	*48*

0 = None-Broken
1 = Broken

TABLE 0.16~~0.16~~ - CONFUSION MATRIX FOR CAST IRON (CALGARY)

Random Forest		**Predicted**		**XGBOOST**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*93*	*467*	**Actual**	**1**	*70*	*490*
	0	*1039*	*12*		**0**	*1016*	*35*

Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*111*	*449*		**1**	*99*	*461*
	0	*1019*	*32*	**Actual**	**0**	*8*	*1043*

0 = Non-Broken
1 = Broken

TABLE 0.17~~0.17~~ - CONFUSION MATRIX FOR ALL MATERIALS – SMOTE METHOD (VANCOUVER)

Random Forest		**Predicted**		**XGBOOST**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*12*	113	**Actual**	**1**	*49*	*76*
	0	*9331*	3192		**0**	*11410*	*1113*

Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**

Actual	**1**	*48*	*77*		**1**	*105*	*20*
	0	*11358*	*1165*	**Actual**	**0**	*12517*	*6*

0 = None-Broken
1 = Broken

TABLE 0.180.18 - CONFUSION MATRIX FOR CAST IRON (VANCOUVER)

Random Forest		**Predicted**		**XGBOOST**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*79*	*14*	**Actual**	**1**	*77*	*16*
	0	*5493*	*2*		**0**	*5490*	*5*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*22*	*71*		**1**	*78*	*15*
	0	*3581*	*1914*	**Actual**	**0**	*5489*	*6*

0 = None-Broken
1 = Broken

TABLE 0.190.19 - CONFUSION MATRIX FOR ALL MATERIALS – SMOTE METHOD (VICTORIA)

Random Forest		**Predicted**		**XGBOOST**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*30*	63	**Actual**	**1**	*45*	*48*
	0	*454*	83		**0**	*511*	*26*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*17*	*76*		**1**	*59*	*34*
	0	*421*	*116*	**Actual**	**0**	*524*	*13*

0 = None-Broken

1 = Broken

TABLE 0.200.20 - CONFUSION MATRIX FOR CAST IRON (VICTORIA)

Random Forest		Predicted		XGBOOST		Predicted	
		0	1			0	1
Actual	1	*30*	*38*	Actual	1	*26*	*42*
	0	*215*	*4*		0	*209*	*10*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*37*	*31*		**1**	*30*	*38*
	0	*213*	*6*	**Actual**	**0**	*213*	*6*

0 = None-Broken
1 = Broken

TABLE 0.210.21 - CONFUSION MATRIX FOR ALL MATERIALS – SMOTE METHOD (HALIFAX)

Random Forest		Predicted		XGBOOST		Predicted	
		0	1			0	1
Actual	1	*71*	298	Actual	1	71	298
	0	*2110*	121		0	2130	101
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	61	308		**1**	105	264
	0	1896	335	**Actual**	**0**	2198	33

0 = None-Broken
1 = Broken

TABLE 0.220.22 - CONFUSION MATRIX FOR CAST IRON (HALIFAX)

Random Forest		Predicted		XGBOOST		Predicted	
		0	1			0	1

Actual	**1**	*53*	*238*	**Actual**	**1**	*56*	*235*
	0	*458*	*30*		**0**	*444*	*44*

Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*52*	*239*		**1**	*46*	*245*
	0	*456*	*32*	**Actual**	**0**	*456*	*32*

0 = None-Broken
1 = Broken

TABLE 0.230.23 - CONFUSION MATRIX FOR ALL MATERIALS – SMOTE METHOD (ST. JOHNS'S)

Random Forest		**Predicted**		**XGBOOST**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*46*	124	**Actual**	**1**	53	117
	0	*1370*	233		**0**	1488	115

Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	48	122		**1**	98	72
	0	1263	340	**Actual**	**0**	1580	23

0 = None-Broken
1 = Broken

TABLE 0.240.24 - CONFUSION MATRIX FOR CAST IRON (ST. JOHNS'S)

Random Forest		**Predicted**		**XGBOOST**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*67*	*69*	**Actual**	**1**	*44*	*92*
	0	*570*	*14*		**0**	*555*	*29*

Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**

Actual	**1**	*37*	*99*		**1**	*75*	*61*
	0	*393*	*191*	**Actual**	**0**	*573*	*11*

0 = None-Broken
1 = Broken

TABLE 0.250~~0.25~~ - CONFUSION MATRIX FOR ALL MATERIALS – SMOTE METHOD (BARRIE)

Random Forest		**Predicted**		**XGBOOST**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*11*	50	**Actual**	**1**	*13*	*48*
	0	*945*	31		**0**	*942*	*34*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*8*	*53*		**1**	*21*	*40*
	0	*872*	*104*	**Actual**	**0**	*967*	*9*

0 = None-Broken
1 = Broken

TABLE 0.260~~0.26~~ - CONFUSION MATRIX FOR CAST IRON (BARRIE)

Random Forest		**Predicted**		**XGBOOST**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*7*	*31*	**Actual**	**1**	*7*	*31*
	0	*31*	*4*		**0**	*29*	*6*
Logistic Regression		**Predicted**		**ANN**		**Predicted**	
		0	**1**			**0**	**1**
Actual	**1**	*6*	*32*		**1**	*8*	*30*
	0	*30*	*5*	**Actual**	**0**	*27*	*8*

0 = None-Broken
1 = Broken

APPENDIX E – REGRESSION RESULTS (ALL CITIES IN DETAIL)

Saskatoon

- ***Age at First Failure***

After cleaning and preparing the dataset for age to the first failure, Saskatoon had 3,332 pipes. This dataset includes a range of input variables such as material, diameter, length, age, lining material, lining status, lining age, and joint type.

Asbestos cement and cast iron are the materials that experienced the highest amount of failure within the network (at least one failure), accounting for 45.95% and 45.63%, respectively. Steel is another material that makes up 5.77% of the dataset. Other materials have not experienced significant failures throughout their life cycles. The given pie chart indicates the percentage of each material that experienced at least one number of failures within the Saskatoon network (Figure 0.1).

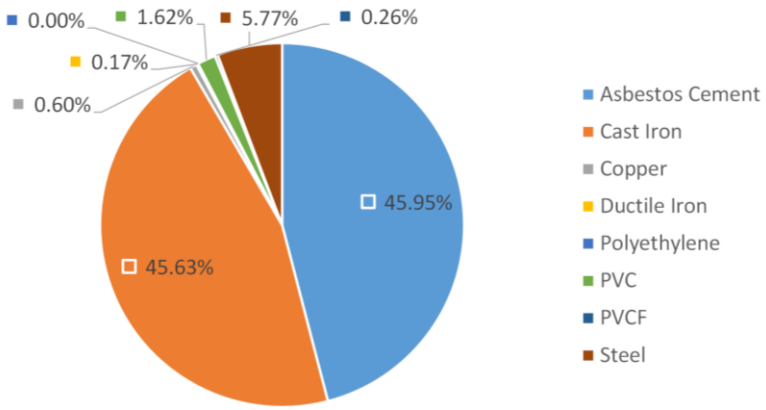


FIGURE 0.1 – PERCENTAGE OF EACH MATERIAL WITHIN REGRESSION ANALYSIS (AGE TO FIRST FAILURE) - SASKATOON

What is essential and should be noted here is the average age to the first failure for various kinds of materials. For example, the average age to the first failure is higher for cast iron pipes compared to others, with a value of 46.64. Additionally, steel pipes and asbestos cement are the following materials, with 40 and 22.62, respectively. PVC pipes, however, have the lowest average age to failure, 10.24. The average age to the first failure for this pipe in the inventory file is around 17 years, including both broken and non-broken pipes. Provided is the box plot showing age distribution to first failure for different materials (Figure 0.2).

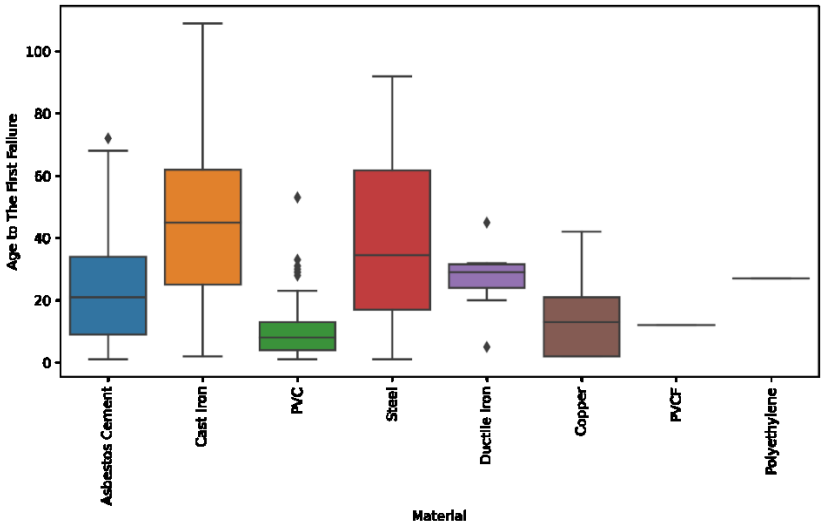


FIGURE 0.2 – DISTRIBUTION OF AGE TO FIRST FAILURE BASED ON TYPE OF MATERIAL (SASKATOON)

Results from regression analysis show moderate satisfaction, although these models should be enhanced for further application. Nevertheless, for AM category, all models represent a somewhat similar result with an RMSE of 17.9. However, random forest indicated a better performance with an MSE of 317.6 and an R-Squared of 0.48. For cast-iron pipes, on the other hand, the random forest had a better performance with an R-Squared of 0.49 and MSE 317.6. Finally, it should be noted that results for random forest and XGBOOST are relatively alike (TABLE 0.1).

TABLE 0.1~~0.1~~ – REGRESSION METRICS – AGE TO FIRST FAILURE (SASKATOON)

Algorithm	RMSE		MSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron	AM	Cast Iron
ElasticNet	*17.9*	*20.3*	*318.8*	*411.4*	*0.47*	*0.45*
Random Forest	*17.8*	*19.7*	*317.6*	*387*	*0.48*	*0.49*
XGBOOST	*17.9*	*19.7*	*321.5*	*387.2*	*0.47*	*0.49*
ANN	*17.9*	*20.0*	*319*	*403.4*	*0.47*	*0.46*

* AM = All Materials

Given scatter plot shows the correlation between actual age to the first failure and predicted age to the first failure based on XGBOOST result. The shape of the chart shows that the model did not perform well for the prediction.

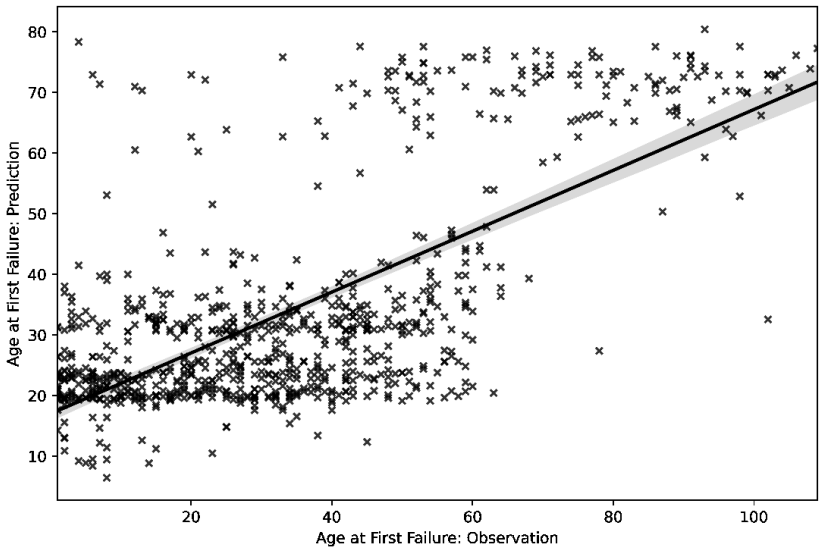


FIGURE 0.3 – REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – XGBOOST) – SASKATOON

- ***Current Rate of Failure***

Furthermore, as previously mentioned current rate of failures was also analyzed separately. The attributes used in this part of the study are material, diameter, length, age, lining material, lining status, lining age, joint type, the previous rate of failures, and the current rate of failures, which is the target of the study. The number of pipes is the same as the previous step. However, here age is the age at the most recent failure. Therefore, the most recent failure could be the nth failure that a pipe experienced. The given figure shows the average current rate of failure for the pipes within the network based on different attributes and plotted versus the age of pipes (Figure 0.4).

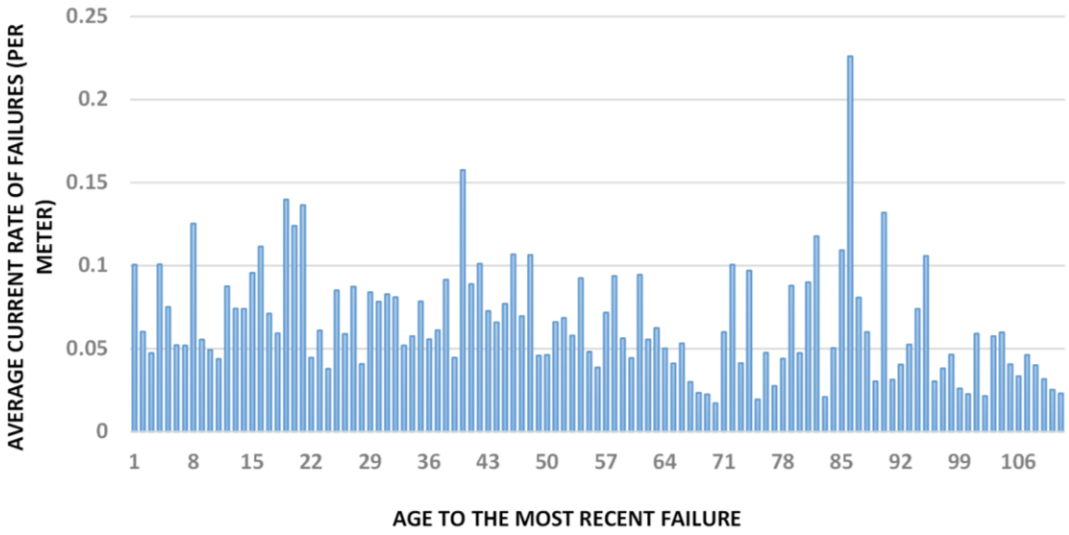


FIGURE 0.4 – AVERAGE OF CURRENT RATE OF FAILURE BASED ON AGE (SASKATOON)

This step indicated that ElasticNet regression does not perform well on the dataset, with an RMSE of 0.13 and an R-Squared of 0.18 (for AM category). Furthermore, for cast iron pipes, also ElasticNet represents a low R-Squared of 0.12. Other models, however, performed satisfactorily. For example, random forest and XBOOST with an RMSE of 0.03 and 0.11 for AM and cast iron categories, respectively, indicated the best performance. However, the result of these algorithms decreased when considering only cast-iron pipes—random forest and XGBOOST both from R-Squared of 0.96 to 0.71. ANN also indicated a relatively desirable score for both categories, RMSE of 0.05 and 0.13 for AM and cast iron categories. The performance of ANN also declined while creating the model for cast iron pipes (TABLE 0.2).

TABLE 0.2~~0.2~~ – REGRESSION METRICS – CURRENT RATE OF FAILURE (SASKATOON)

Algorithm	RMSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron
ElasticNet	*0.13*	*0.20*	*0.18*	*0.12*
Random Forest	*0.03*	*0.11*	*0.96*	*0.71*
XGBOOST	*0.03*	*0.12*	*0.96*	*0.71*
ANN	*0.05*	*0.13*	*0.85*	*0.62*

* AM = All Materials

The given plot depicts the correlation between the actual current rate of failures and the predicted current rate of failures created based on the random forest result (Figure 0.5).

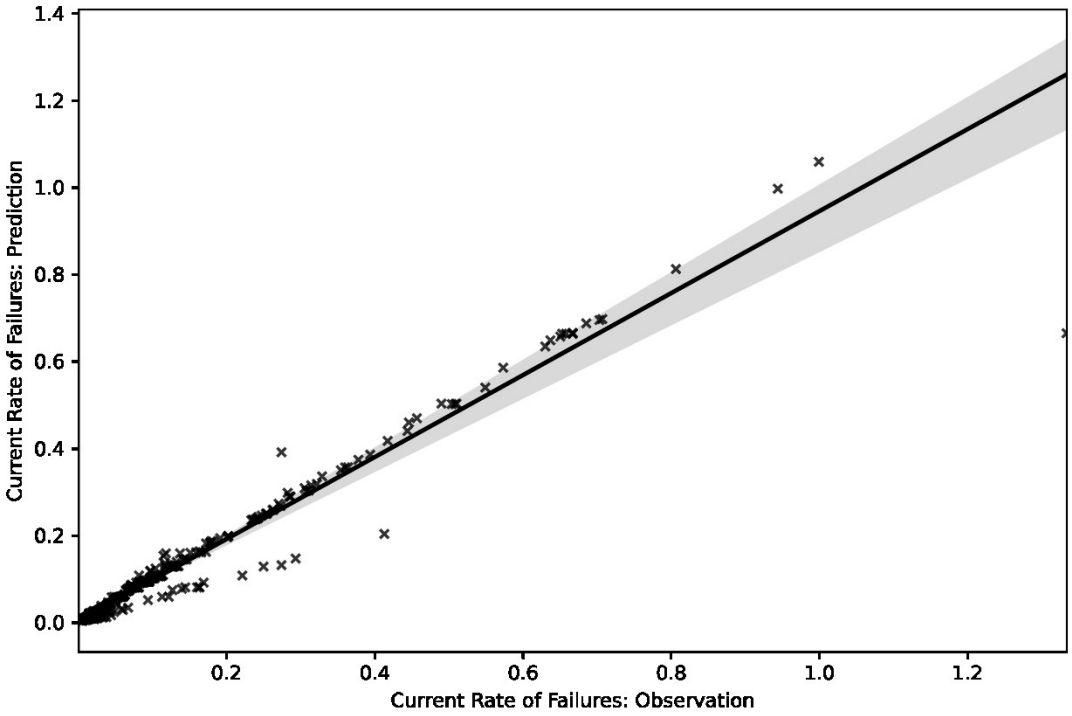


FIGURE 0.5 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – RANDOM FOREST) - SASKATOON

Winnipeg

- ***Age at First Failure***

With one of the largest datasets, this network includes 6,913 pipes that experienced at least one number of failures. This dataset consists of a range of input variables such as material, diameter, joint type, length, coating materials, and age of pipe at first failure. From the given chart, it is clear that cast iron pipe accounts for 65.34% of this file, with the highest number of failures (Figure 0.6). Asbestos cement is another type of material with almost 29.64% contribution to the failures. Finally, PVC and ductile iron are the following materials with 2.63% and 2.31%, respectively.

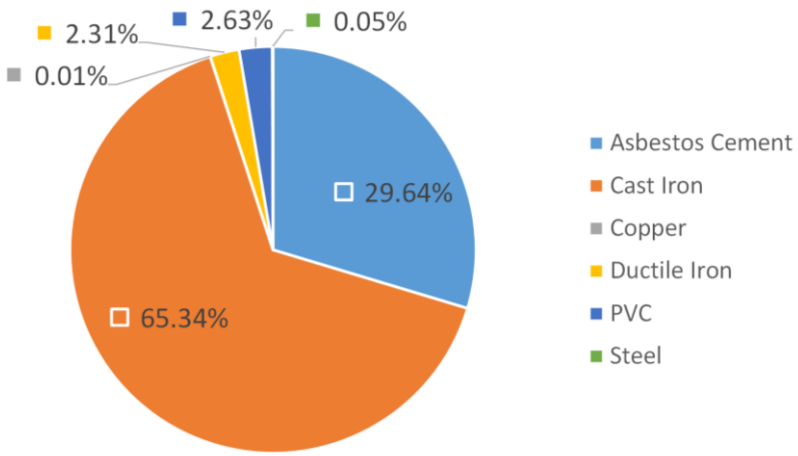


FIGURE 0.6 - PERCENTAGE OF EACH MATERIAL WITHIN REGRESSION ANALYSIS (AGE TO FIRST FAILURE) – WINNIPEG

The given chart shows the distribution of age to the first failure based on different materials within the network (Figure 0.7). From the information below, cast iron with an average of 59.11 has the highest age to the first failure, followed by steel pipes with 52.71. Asbestos cement and ductile iron are the following materials, with average age to failure of 29.81 and 26.07, in successive. PVC pipes seem to have experienced the first failure during their early life cycle stage with a value of 13.86. This pipe is more prone to failure when it is young compared to others. The susceptibility of PVC pipes to failure during the early stage could have different causes based on the site of the study, which requires more investigations.

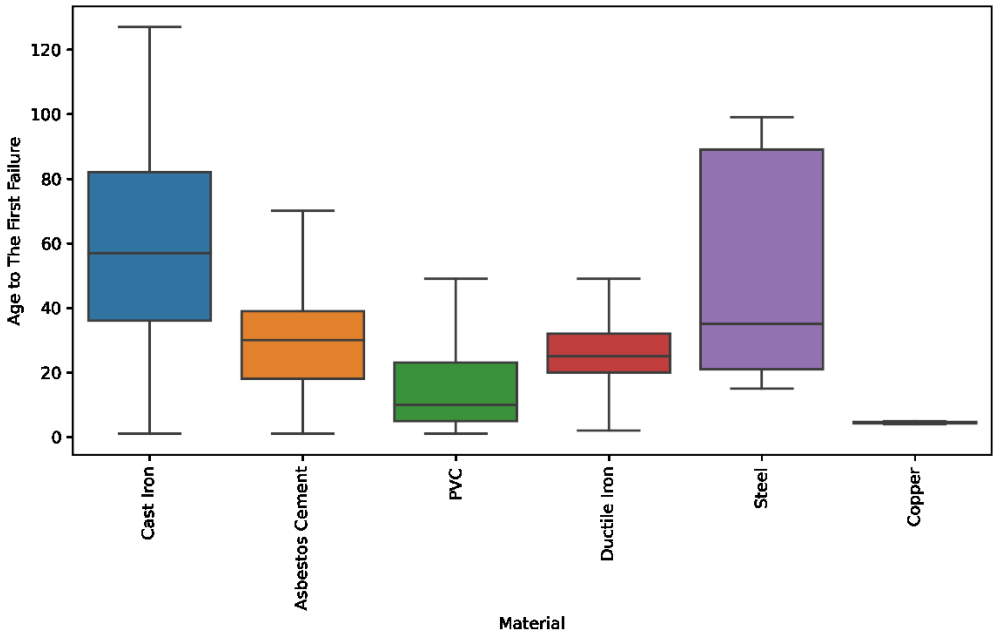


FIGURE 0.7 - DISTRIBUTION OF AGE TO FIRST FAILURE BASED ON TYPE OF MATERIAL (WINNIPEG)

Results from the analytical process indicate an unsatisfactory performance for all algorithms, considering age to the first failure (TABLE 0.3). Nonetheless, random forest and XGBOOST models performed better compared to EassticNet and ANN. For AM category, these two algorithms showed a similar accuracy with an RMSE score of 19.5 and an R-squared of 0.47.

For cast-iron pipes also random forest and XGBOOST with an RMSE of 21.6 had the best performance. It should be noted that the accuracy of the models declined when a homogenous group of pipe (Cast Iron) was analyzed. For instance, in this case, for the random forest, RMSE increased from 19.5 to 21.6 shows that not a similar group of pipes necessarily enhances the results. Results for other models can be found in the table.

TABLE 0.3~~0.3~~ - REGRESSION METRICS (WINNIPEG)

Algorithm	RMSE		MSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron	AM	Cast Iron
ElasticNet	*19.8*	*22*	*393*	*479.4*	*0.45*	*0.3*
Random Forest	*19.5*	*21.6*	*379.5*	*467.5*	*0.47*	*0.31*
XGBOOST	*19.5*	*21.6*	*382.3*	*468.1*	*0.47*	*0.31*
ANN	*21.9*	*21.9*	*477.7*	*481.1*	*0.34*	*0.3*

* AM = All Materials

The provided regression plot from the seaborn library in python shows that XGBOOST, one of the best models for this network, could not find an appropriate pattern for this dataset (Figure 0.8). As shown in the table, R-Squared for XGBOOST is 0.47, which is relatively low for making a prediction (TABLE 0.3).

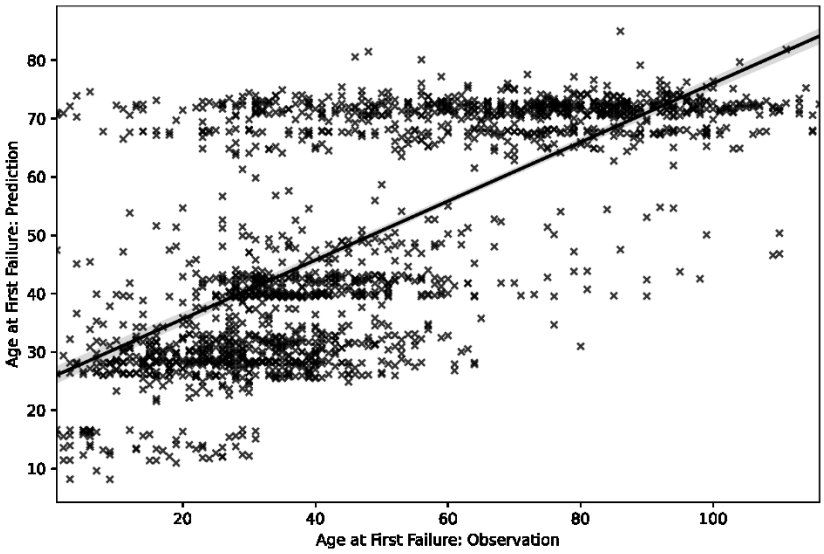


FIGURE 0.8 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – XGBOOST) - WINNIPEG

- ***Current Rate of Failure***

The current rate of failure was also analyzed for this utility. In this step, material, diameter, joint type, length, coating material, age, the previous rate of failure (PreviousRoF), and, more importantly, the current rate of failure (CurrentRoF) were utilized as input variables. This network shows that the average rate of failure is higher during the early stage and final stage of the life cycle, compared to the Bath-Tub curve mentioned in the previous chapters. The given chart shows the distribution of the average of the current rate of failures based on the age to the most recent failure (Figure 0.9).

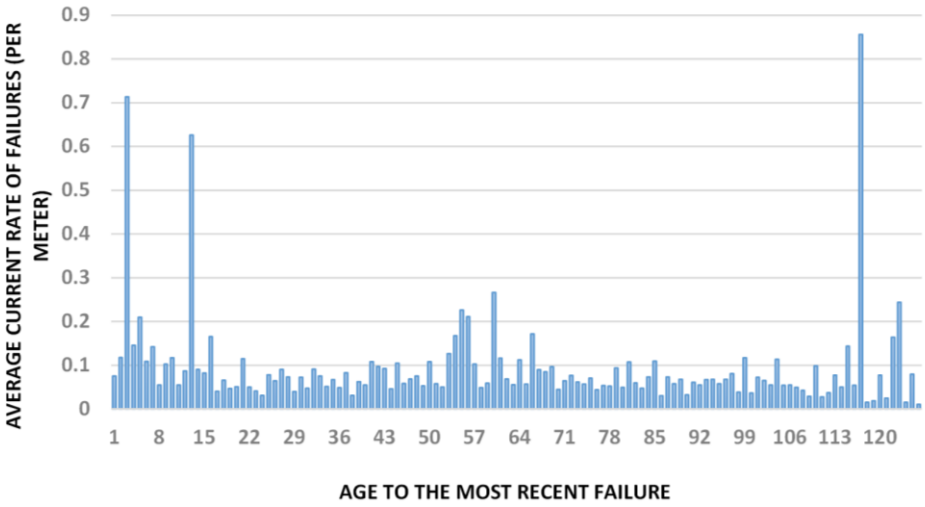


FIGURE 0.9 - AVERAGE OF CURRENT RATE OF FAILURE BASED ON AGE (WINNIPEG)

Considering the RMSE score and R-Squared, XGBOOST indicated the best performance for both categories. For AM category, an RMSE score of 0.05 and R-Squared of 0.99 was the accuracy of

the XGBOOST model. ANN model, however, did not perform well, with an RMSE score of 0.70. Therefore, this model was not able to calculate R-Square for AM category (TABLE 0.4).

For cast-iron pipes also with an R-Squared of 0.97, XGBOOST represented the best performance. On the other hand, ANN and random forest indicated a relatively good performance for the cast iron category with 0.84 and 0.91 for R-Squared, respectively. It is worth mentioning that some of these algorithms are significantly sensitive to hyperparameter tuning, meaning that the accuracy of ANN, for instance, may or may not increase with only changing the number of nodes or hidden layers. Therefore, playing around and finding the best hyperparameters for the neural network may require more time, and also it is an expensive process.

TABLE 0.4~~0.4~~ – REGRESSION METRICS – CURRENT RATE OF FAILURE (WINNIPEG)

Algorithm	RMSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron
ElasticNet	*0.44*	*0.34*	*0.03*	*0.05*
Random Forest	*0.14*	*0.05*	*0.91*	*0.91*
XGBOOST	*0.05*	*0.06*	*0.99*	*0.97*
ANN	*0.70*	*0.14*	*N.A*	*0.84*

* AM = All Materials

The given regression graph plotted the observed rate of failure versus the predicted rate of failure, based on the XGBOOST results. It can be seen that the model was able to find a good pattern for this dataset. However, to put this model in practice, more investigation is required to ensure the reliability of the performance (Figure 0.10).

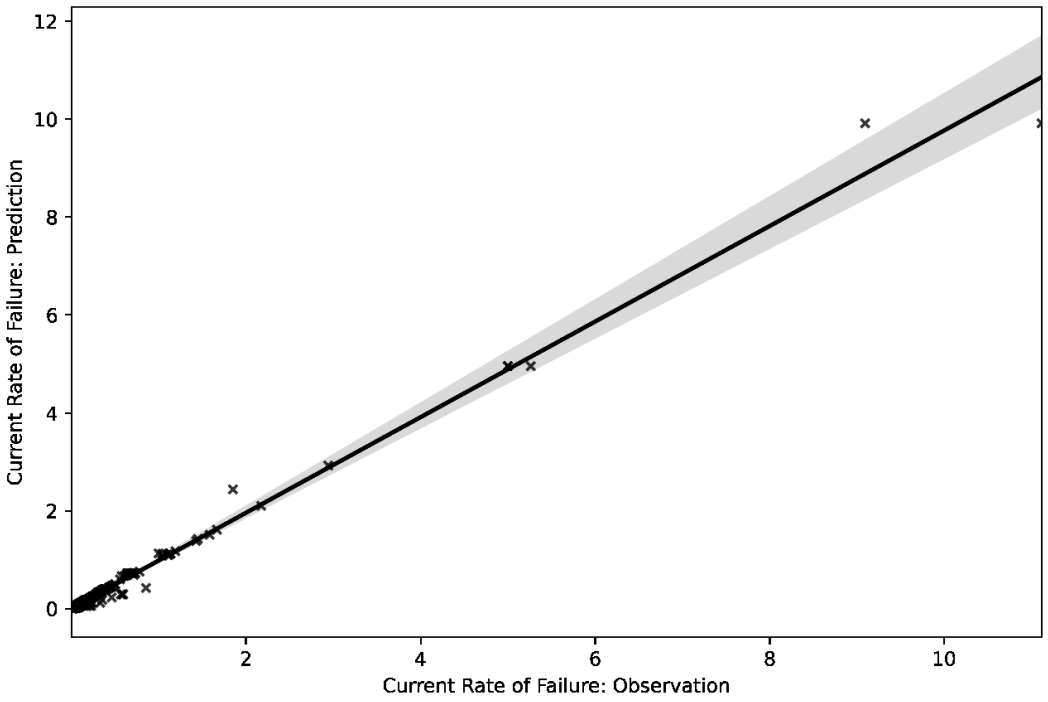


FIGURE 0.10 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – XGBOOST) - WINNIPEG

Kitchener

- ***Age at First Failure***

For regression analysis, this network consists of 917 unique pipes with at least one number of failures, including different attributes such as anode status, material, lining material, diameter, lining status, length, lining age, and age.

Once more, cast iron is the most frequent type of material within the prepared file, with 69.03% of the total. Ductile iron follows this material with a little more than a quarter of total pipes, 27.70%, and PVC only accounts for 2.29% of this network. The percentage of all materials can be seen in the given pie chart in more detail (Figure 0.11). Copper has a minor contribution to failures in Kitchener, with only 0.11% of overall failures. Only pipe with at least one failure was considered in this part of the study, and only the first failure was used for this dataset.

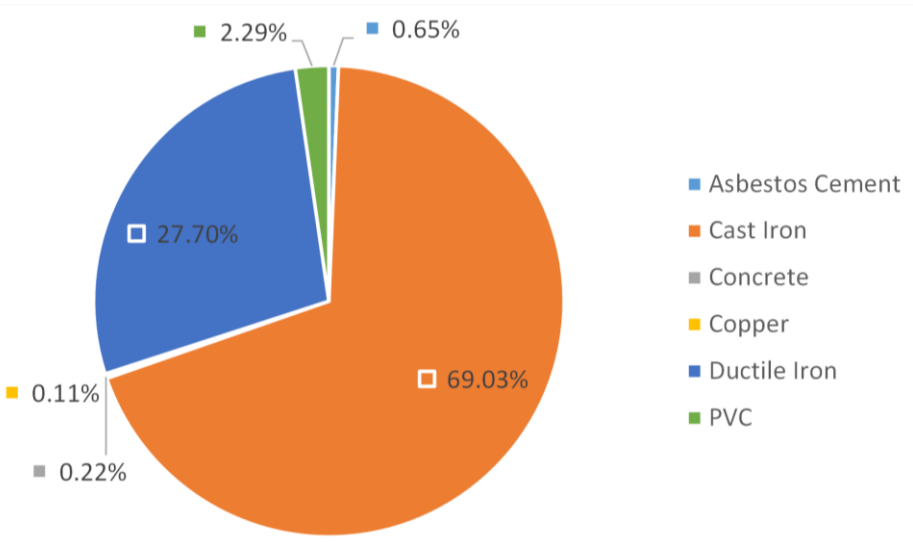


FIGURE 0.11 - PERCENTAGE OF EACH MATERIAL WITHIN REGRESSION ANALYSIS (AGE TO FIRST FAILURE) - KITCHENER

In terms of average age to the first failure, cast iron shows a higher value, 49.61 years. This material in the majority of utilities shows the highest average age to the first failure. Cast iron then is followed by asbestos cement and ductile iron with an average of 42.17 and 34.54, respectively. Copper, concrete, and PVC are other materials with a lower age average to the first failure. PVC has the lowest average, which is approximately 11.19 years that is relatively young. The given chart compares the average age to the first failure for all materials within this network (Figure 0.12), showing age distribution.

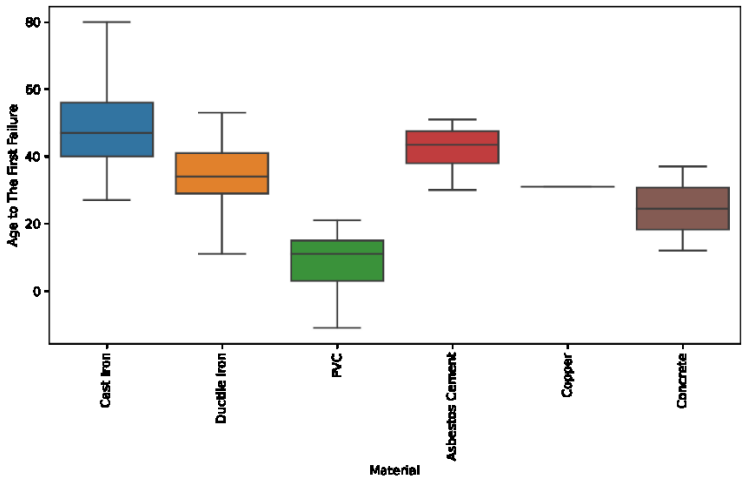


FIGURE 0.12 - DISTRIBUTION OF AGE TO FIRST FAILURE BASED ON TYPE OF MATERIAL (KITCHENER)

Comparing the results given in the chart below shows that the random forest algorithm had the best performance for predicting age to the first failure, with an RMSE score of 11.3 and R-Squared of 0.39, for AM category (TABLE 0.5). Moreover, the Random forest indicated a better

accuracy for the cast iron group with an RMSE of 13.0 and R-Squared of 0.30. It should be noted that XGBOOST showed the lowest accuracy for AM category, with an R-Squared of 0.34 and an RMSE of 11.7. The ANN, however, indicated the worst performance of cast iron pipes.

TABLE 0.5~~0.5~~ - REGRESSION METRICS (KITCHENER)

Algorithm	RMSE		MSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron	AM	Cast Iron
ElasticNet	*11.3*	*13.2*	*128.5*	*175.4*	*0.38*	*0.28*
Random Forest	*11.3*	*13.0*	*127.5*	*168.5*	*0.39*	*0.30*
XGBOOST	*11.7*	*13.9*	*136.7*	*192.7*	*0.34*	*0.21*
ANN	*11.4*	*14*	*128.9*	*196.1*	*0.38*	*0.19*

* AM = All Materials

The correlation between the observed and predicted age to the first failure can be seen in the given graph (Figure 0.13).

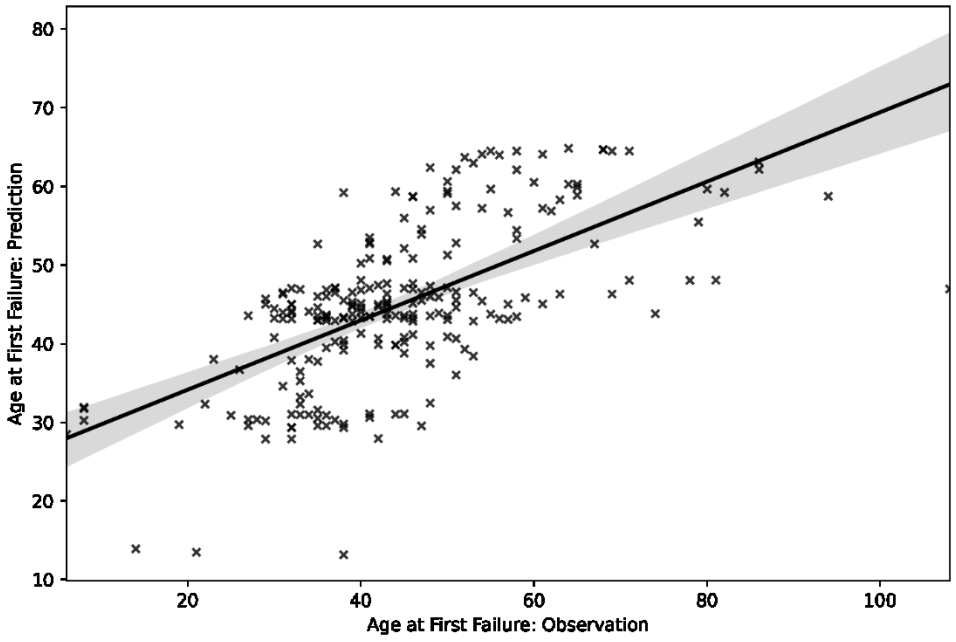


FIGURE 0.13 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – RANDOM FOREST) – KITCHENER

- ***Current Rate of Failure***

After cleaning and preparing the dataset for the current rate of failures, the following input variables remained in the analysis: anode status, material, lining material, diameter, lining status, age, lining age, the current rate of failures, and previous rate of failures.

As can be seen in the given chart, the trend for the average current rate of failure is similar to that of Winnipeg (Figure 0.14). Apparently, the rate of failure is higher during 1st stage and final stage of pipes within the network. This indicates that pipes are more prone to failure in these two stages.

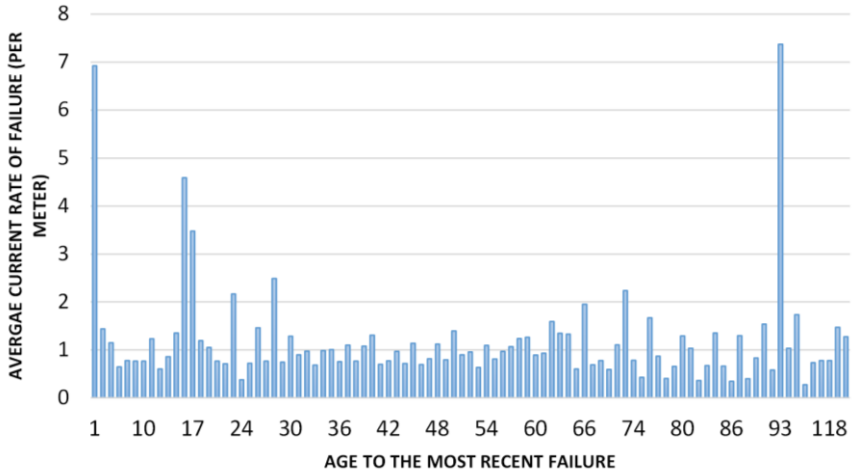


FIGURE 0.14 - AVERAGE OF CURRENT RATE OF FAILURE BASED ON AGE (KITCHENER)

The given table below compares the result of different models (TABLE 0.6). For example, the ElasticNent model did not show a satisfactory performance with a low R-Squared, 0.29. However, XGBOOST and ANN indicated a relatively high accuracy for AM category, with an R-Squared of 0.91.

Furthermore, for cast-iron pipes, ANN performed better than other models with an RMSE of 0.32 and an R-Squared of 0.73. Finally, although not being the best model, random forest indicated a relatively similar performance to ANN and XGBOOST. Therefore, based on the XGBOOST results, a regression plot was prepared to compare the predicted and actual values (Figure 0.15).

TABLE 0.6~~0.6~~ – REGRESSION METRICS – CURRENT RATE OF FAILURE (KITCHENER)

Algorithm	RMSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron
ElasticNet	*1.07*	*0.60*	*0.29*	*0.05*
Random Forest	*0.52*	*0.34*	*0.83*	*0.70*
XGBOOST	*0.38*	*0.33*	*0.91*	*0.71*
ANN	*0.40*	*0.32*	*0.91*	*0.73*

* AM = All Materials

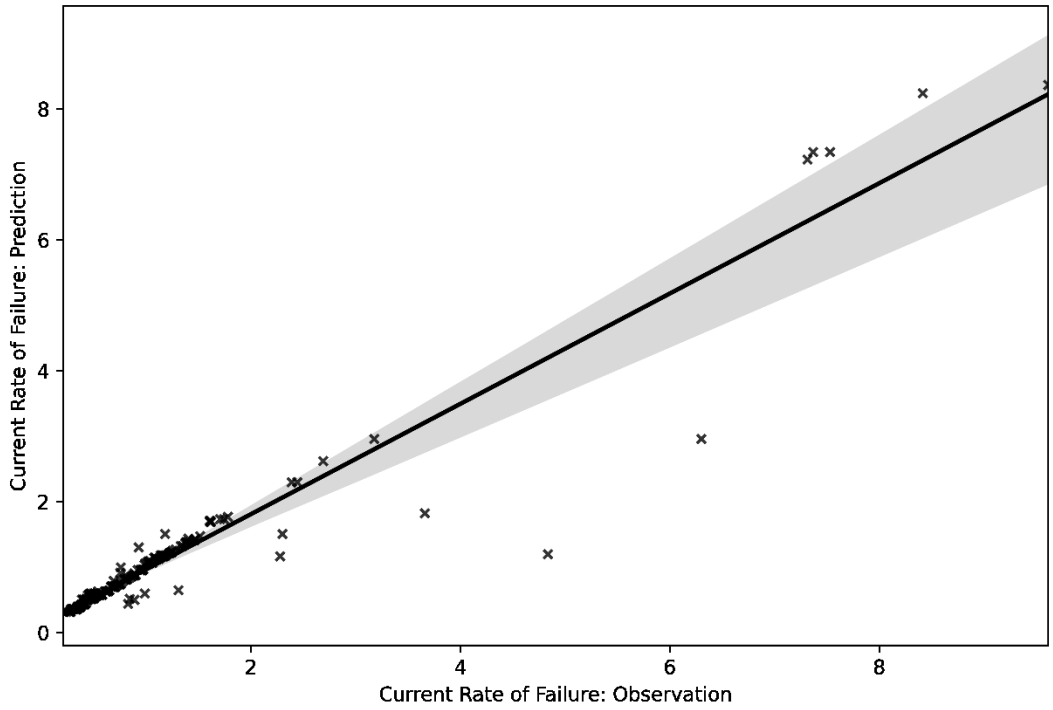


FIGURE 0.15 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – XGBOOST) - KITCHENER

Markham

- ***Age at First Failure***

Markham city consists of 591 unique pipes with at least one failure. This utility includes different features, including material, diameter, length, lining status, pipe depth, protection status, protection age, lining age, and the target variable, age.

Ductile iron is the most frequent material within this network, accounting for 56.18% of the entire dataset. Cast iron follows ductile iron with a 32.99% contribution to total failures. PVC with only 8.63% is another frequent material within this network. The given pie charts provide more information regarding the percentage of each material (Figure 0.16).

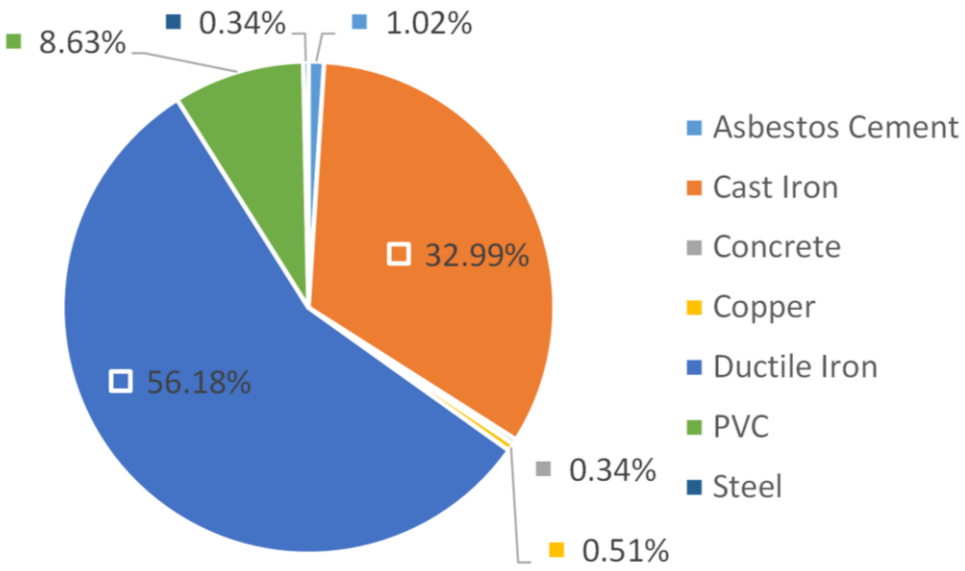


FIGURE 0.16 - PERCENTAGE OF EACH MATERIAL WITHIN REGRESSION ANALYSIS (AGE TO FIRST FAILURE) – MARKHAM

In this network, asbestos cement and concrete pipes have the highest average age to the first failure, with the value of 38.17 and 26, respectively. Furthermore, cast iron and steel pipes with the age of almost 22 are the following materials. On the other hand, PVC is the material with the minimum average age to the first failure, indicating the material's vulnerability during younger age. The given box plot indicates the distribution of age based on different materials in Markham (Figure 0.17).

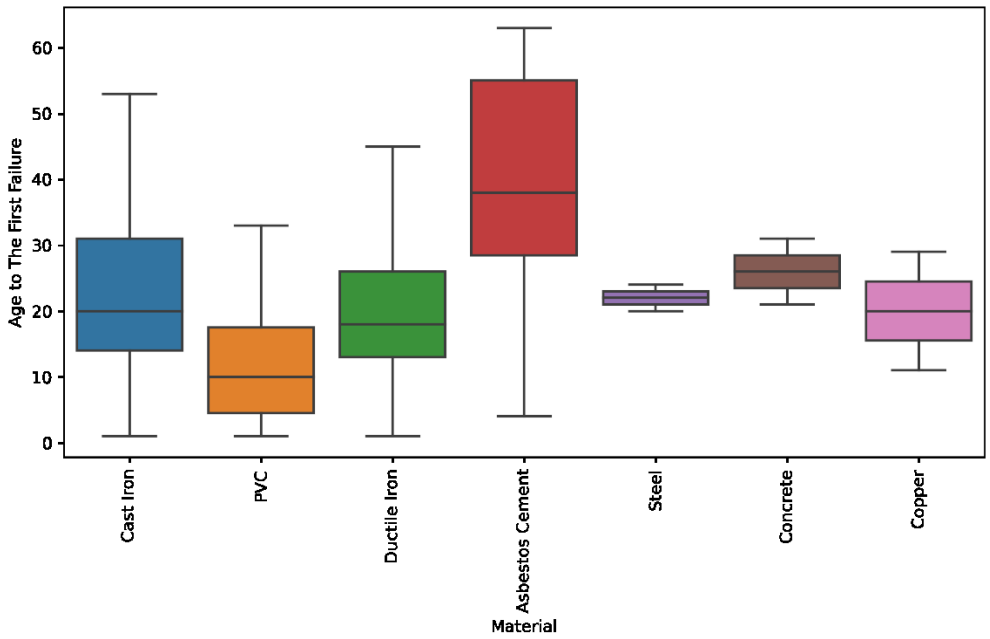


FIGURE 0.17 - DISTRIBUTION OF AGE TO FIRST FAILURE BASED ON TYPE OF MATERIAL (MARKHAM)

Analyzing the regression results shows the low accuracy for all models. For instance, the random forest is the best model for AM category, which provides prediction only with an RMSE score of 11.1 and the R-Squared of 0.17, which is significantly low. For cast-iron pipes, on the other hand, ElasticNet regression has the RMSE of 11.8, which is more desirable than other models. It should be noted that finding an efficient regression model for this city requires more investigation. The given table below represents more information about the accuracy of these algorithms (TABLE 0.7).

TABLE 0.7~~0.7~~ - REGRESSION METRICS (MARKHAM)

Algorithm	RMSE		MSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron	AM	Cast Iron
ElasticNet	*11.2*	*11.8*	*124.9*	*138.8*	*0.15*	*N.A*
Random Forest	*11.1*	*12.3*	*122.1*	*151.1*	*0.17*	*N.A*
XGBOOST	*11.6*	*13.2*	*135.6*	*172.9*	*0.08*	*N.A*
ANN	*12.3*	*11.9*	*151.1*	*140.5*	*N.A*	*N.A*

* AM = All Materials

Provided regression plot indicates how well the random forest model fits the dataset by comparing actual and predicted values (Figure 0.18).

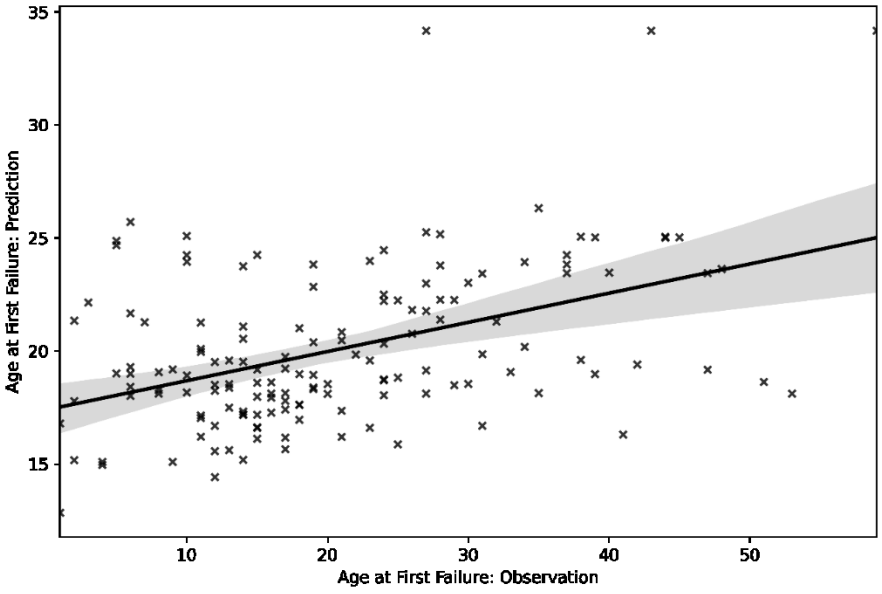


FIGURE 0.18 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – RANDOM FOREST) – MARKHAM

- ***Current Rate of Failure***

From the given chat, it is clear that there is a fluctuation for different ages, and no specific pattern can be seen (Figure 0.19). Nonetheless, pipes with ages around 5, 21, 36, and 58 have experienced more failures than others.

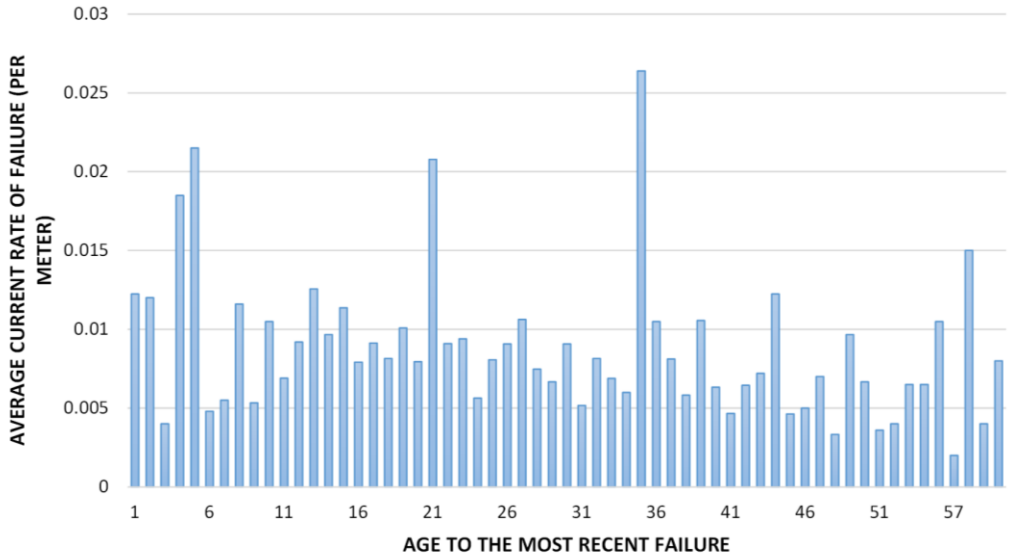


FIGURE 0.19 - AVERAGE OF CURRENT RATE OF FAILURE BASED ON AGE (MARKHAM)

Once more, the extracted results reveal the powerfulness of tree-based algorithms. For example, random forest and XGBOOST performed better than ELasticNet and ANN algorithms (TABLE 0.8). The RMSE for these two models is the same, which is 0.003 for AM category. However, in terms of R-Squared, XGBOOST performed better than random forest, with a score of 0.87. The ANN was the worst model for this network, with an RMSE of 0.02. This algorithm did not provide a logical R-Squared in this step.

It is essential to find out how well a regression model fits the dataset. In order to do so, a regression plot has been prepared, which compares the actual value and the predicted current rate of failures within Markham`s network (Figure 0.20).

TABLE 0.8~~0.8~~ – REGRESSION METRICS – CURRENT RATE OF FAILURE (MARKHAM)

Algorithm	RMSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron
ElasticNet	*0.008*	*0.003*	*N.A*	*0.38*
Random Forest	*0.003*	*0.004*	*0.81*	*N.A*
XGBOOST	*0.003*	*0.003*	*0.87*	*0.43*

ANN	*0.02*	*0.02*	*N.A*	*N.A*

* AM = All Materials

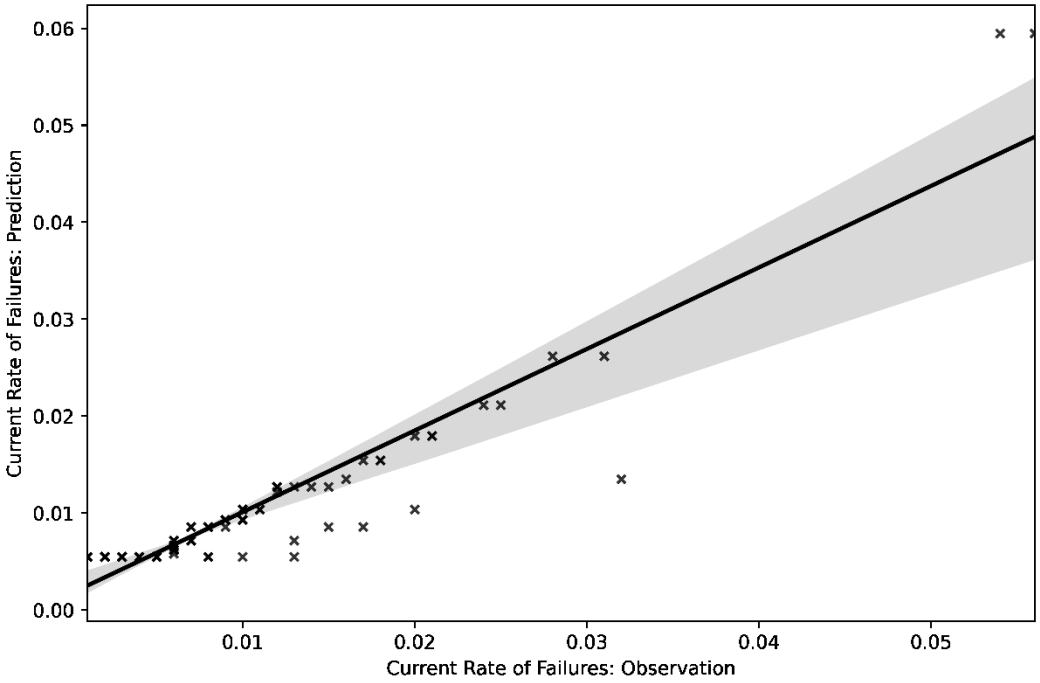


FIGURE 0.20 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – XGBOOST) - MARKHAM

Waterloo

- ***Age at First Failure***

Waterloo is another network that has been analyzed in terms of age to first failure and the current failure rate. This network comprises 432 unique pipes for this step of the analysis, and for this part, it includes diameter, material, lining age, lining status, lining material, length, and age to the first failure. About 81.71% of the total pipes belong to the cast iron material, followed by ductile iron, which accounts for 15.28% of the dataset. PVC and HDPE have a small proportion with 2.78% and 0.23%, respectively (Figure 0.21).

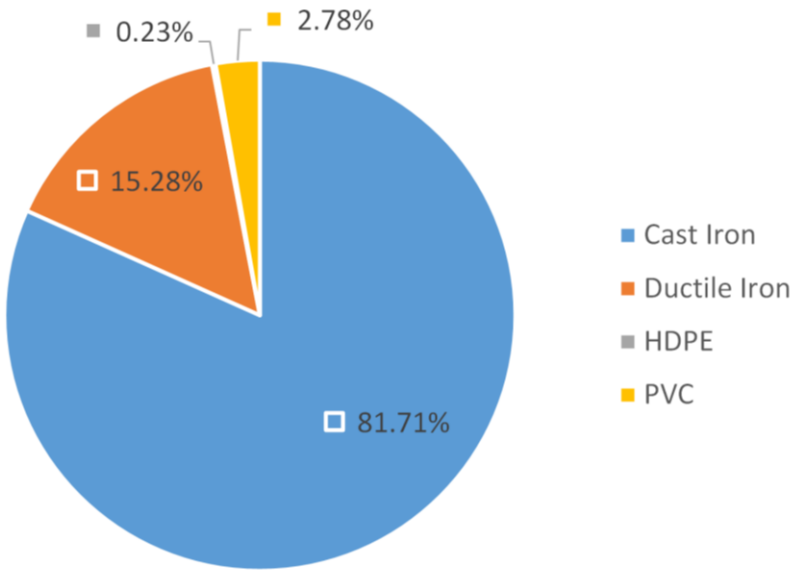


FIGURE 0.21 - PERCENTAGE OF EACH MATERIAL WITHIN REGRESSION ANALYSIS (AGE TO FIRST FAILURE) – WATERLOO

For this network, cast iron has the highest average age to the first failure, and HDPE has the minimum average age, with 49.3 and 18 years, respectively. Ductile iron and PVC are the other materials existing in this network, and the distribution of each can be seen within the given bar chart (Figure 0.22).

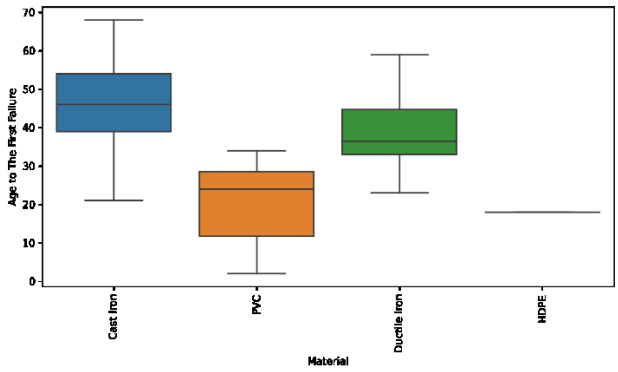


FIGURE 0.22 - DISTRIBUTION OF AGE TO FIRST FAILURE BASED ON TYPE OF MATERIAL (WATERLOO)

Comparing all results shows that XGBOOST has the best performance for AM category with an RMSE of 9.8 and R-Squared of 0.42. On the other for cast iron pipes, ANN showed a better accuracy compared to others. This algorithm with an RMSE of 10.5 and an R-Squared of 0.34 was the best model. Other values can be found within the given table (TABLE 0.9). The following regression plot shows how accurate XGBOOST is by comparing actual and predicted values (Figure 0.23).

TABLE 0.90.9 - REGRESSION METRICS (WATERLOO)

Algorithm	RMSE		MSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron	AM	Cast Iron
ElasticNet	*11.8*	*12.8*	*138.9*	*164.5*	*0.16*	*0.02*
Random Forest	*10.6*	*11.4*	*111.8*	*130.8*	*0.32*	*0.22*
XGBOOST	*9.8*	*11.9*	*96.4*	*143.2*	*0.42*	*0.15*
ANN	*11.8*	*10.5*	*138*	*110.6*	*0.16*	*0.34*

* AM = All Materials

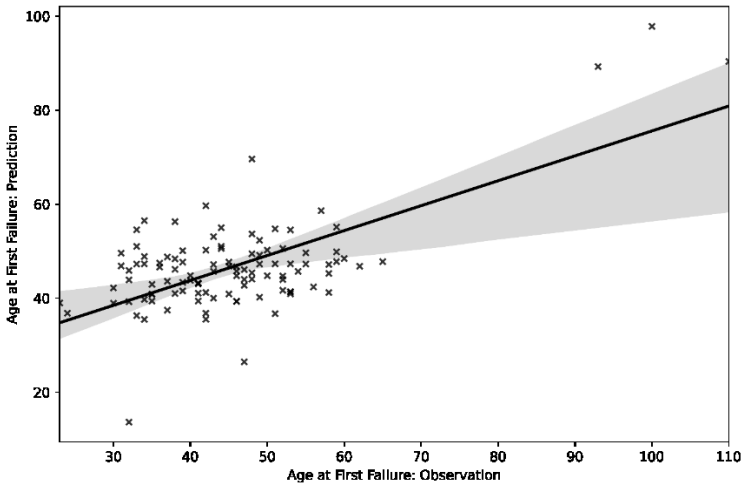


FIGURE 0.23 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – XGBOOST) – WATERLOO

- ***Current Rate of Failure***

Additionally, several input variables were employed for the current rate of failure, containing diameter, material, lining status, lining material, length, lining age, age at most recent failure, the previous rate of failures, and the current rate of failures. It should be noted that the average current rate of failure has experienced a fluctuation depending on different ages. From the given chart, it can be seen that from installation, the current rate of failure increased until 36 (Figure 0.24). This trend then changed, and the average current rate of failures declined. However, there is a significant peak for pipes over 107 related to the wear-out phase of the bathtub curve.

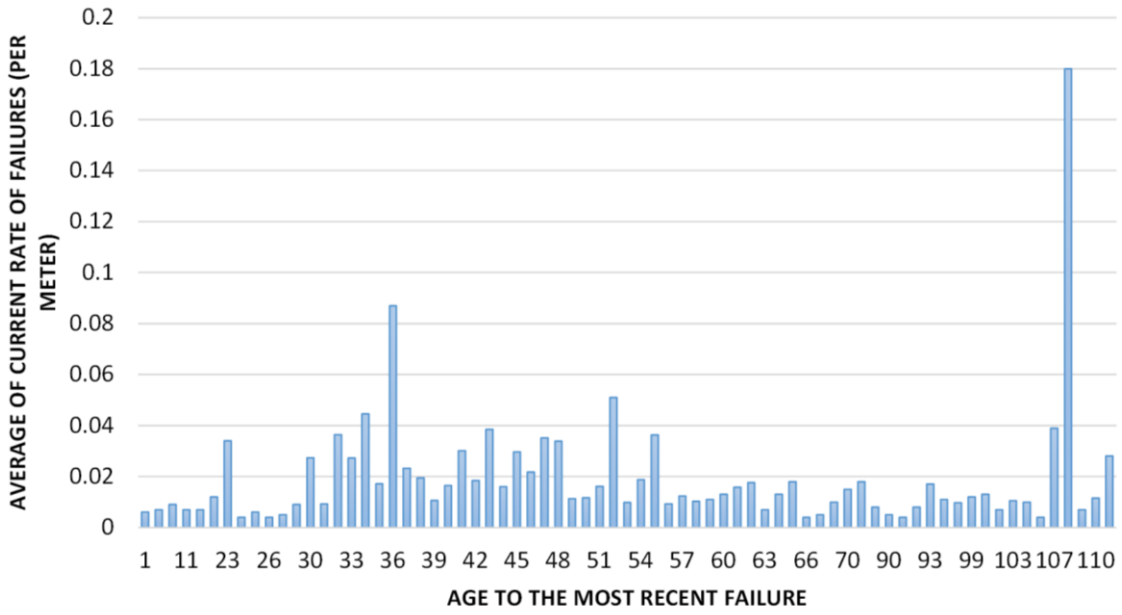


FIGURE 0.24 - AVERAGE OF CURRENT RATE OF FAILURE BASED ON AGE (WATERLOO)

For AM category, once more, XGBOOST depicted a better performance than other models, with RMSE and R-Squared of 0.03 and 0.78, respectively. On the other hand, ElasticNet was the weakest model for this category. Moreover, the random forest and ANN results were not desirable models with R-Squared of 0.58 and 0.44, in successive (TABLE 0.10).

It should be noted that, in this utility with partitioning the dataset based on material, the overall performance of all models increased relatively. For instance, the random forest`s R-Squared increased from 0.58 for AM category to 0.73 for cast iron pipes, indicating that the partitioning step may increase the accuracy in some cases.

TABLE 0.100.10 – REGRESSION METRICS – CURRENT RATE OF FAILURE (WATERLOO)

Algorithm	RMSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron
ElasticNet	*0.06*	*0.05*	*0.12*	*0.16*
Random Forest	*0.04*	*0.03*	*0.58*	*0.73*
XGBOOST	*0.03*	*0.02*	*0.78*	*0.88*
ANN	*0.05*	*0.03*	*0.44*	*0.66*

* AM = All Materials

Similar to other utilities, a regression plot was prepared for the best model, the XGBOOST algorithm, in this case (Figure 0.25).

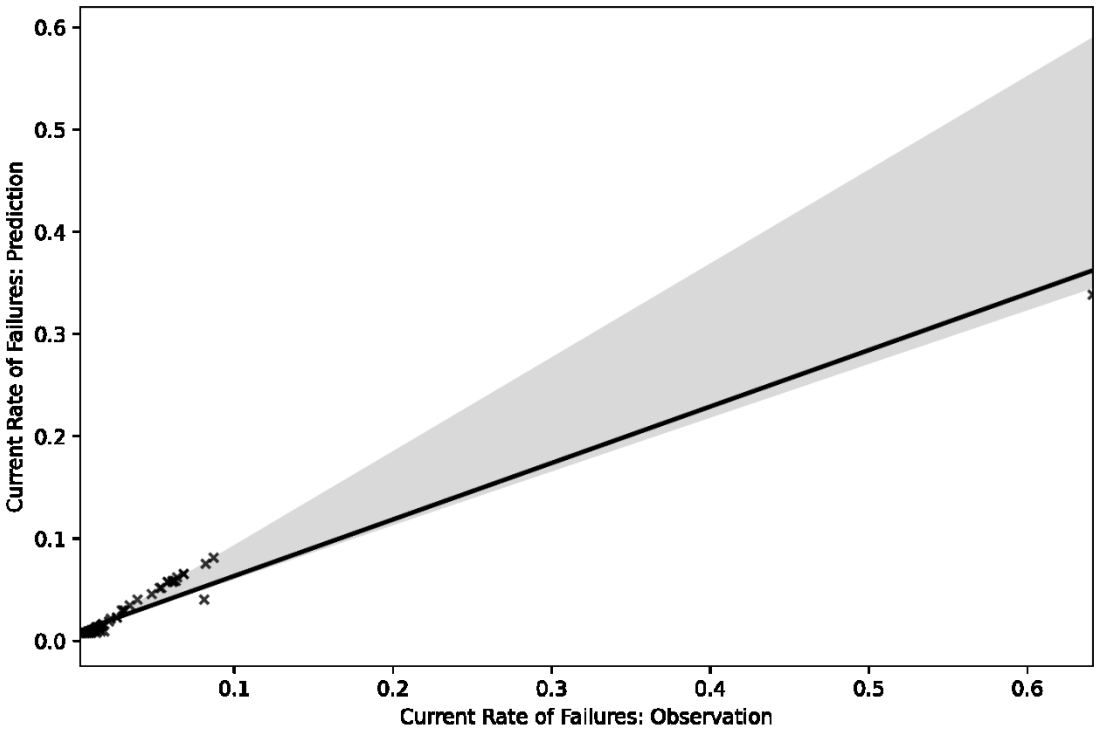


FIGURE 0.25 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – XGBOOST) - WATERLOO

Region of Waterloo

- ***Age at First Failure***

For this utility, after the cleaning process, several input variables remained for regression analysis. These attributes include material, diameter, length, lining status, lining material, lining age, and age to the first failure. This network consists of merely 92 unique pipes that experienced at least one failure. The significantly low number of broken pipes made both classification and regression analysis more challenging.

Ductile iron accounts for 38.04% of these pipes, and cast iron makes up 34.78% of the total. PVC, asbestos cement, and concrete are the other materials with 14.13%, 6.52%, and 6.52% contribution to total failures, respectively (Figure 0.26).

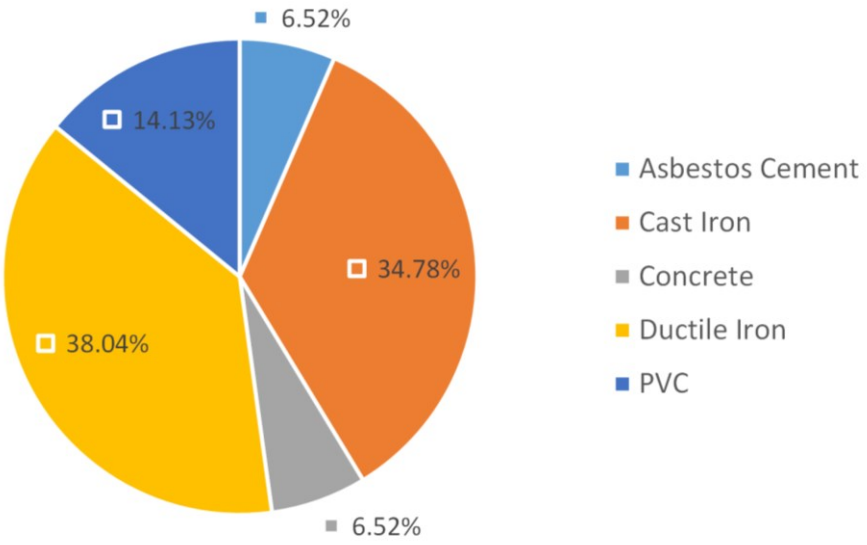


FIGURE 0.26 - PERCENTAGE OF EACH MATERIAL WITHIN REGRESSION ANALYSIS (AGE TO FIRST FAILURE) – REGION OF WATERLOO

Like most utilities, the cast iron pipe has the highest average years to the first failure, 59. This number for asbestos cement and concrete is 43.83 and 33.17, in successive. PVC pipes, however, experienced failures in the early stage of their life cycles, with an average of 14.77 years to the first failure (Figure 0.27).

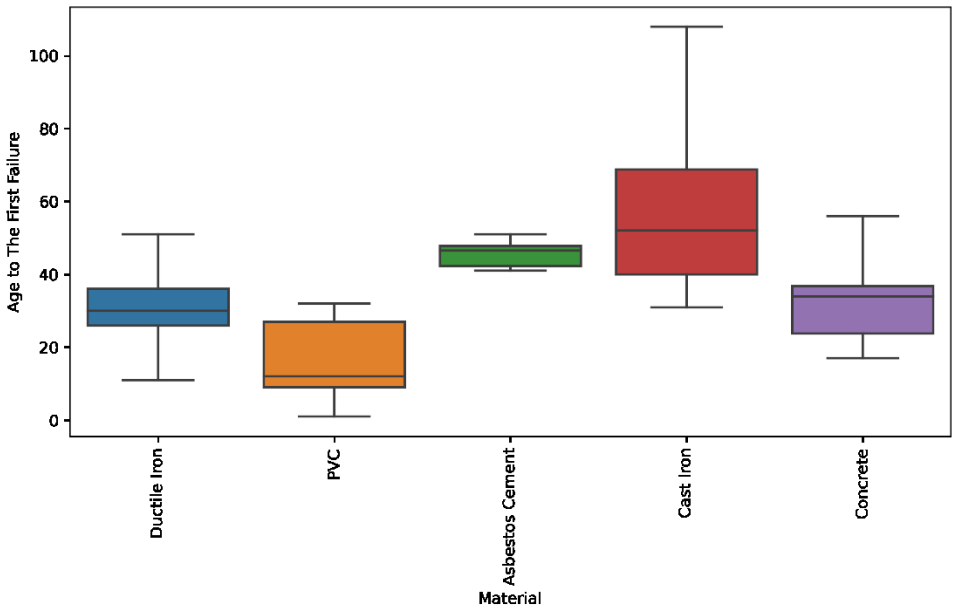


FIGURE 0.27 - DISTRIBUTION OF AGE TO FIRST FAILURE BASED ON TYPE OF MATERIAL (REGION OF WATERLOO)

It should be noted that since the number of cast iron pipes was not sufficient, only AM category, which includes all material, was analyzed for the Region of Waterloo. In this section, ANN algorithm showed the worst accuracy with an RMSE of 22.4 and an R-Squared of 0.22. The random forest model, however, depicted a relatively good score. With the RMSE of 14.5 and the R-Squared of 0.68, this algorithm was the most efficient one. Results for other algorithms are provided in the following table (TABLE 0.11). Given regression plot also provides more insight regarding the accuracy of the random forest as the best model for the Region of Waterloo (Figure 0.28).

TABLE 0.110.11 - REGRESSION METRICS (WATERLOO)

Algorithm	RMSE		MSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron	AM	Cast Iron
ElasticNet	*21.5*	-	*461.2*	-	*0.29*	-
Random Forest	*14.5*	-	*209.2*	-	*0.68*	-
XGBOOST	*17.4*	-	*303.6*	-	*0.53*	-
ANN	*22.4*	-	*503.1*	-	*0.22*	-

* AM = All Materials

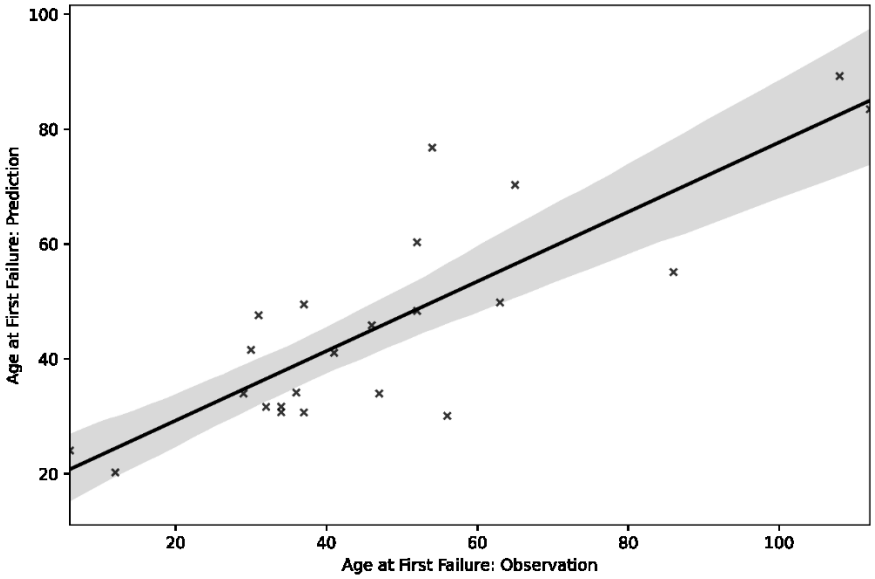


FIGURE 0.28 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – RANDOM FOREST) – REGION OF WATERLOO

- ***Current Rate of Failure***

Furthermore, this utility was also analyzed for the current rate of failure. Length, material, diameter, lining status, lining material, lining age, age at the most recent failure, the previous rate of failure, and the current failure rate are the input variables for this step. Since this utility is a young network, the average current rate of failure is higher for younger pipes, and this can be seen from the given graph. The graph shows that pipes aged around 20 experienced more failures than others (Figure 0.29).

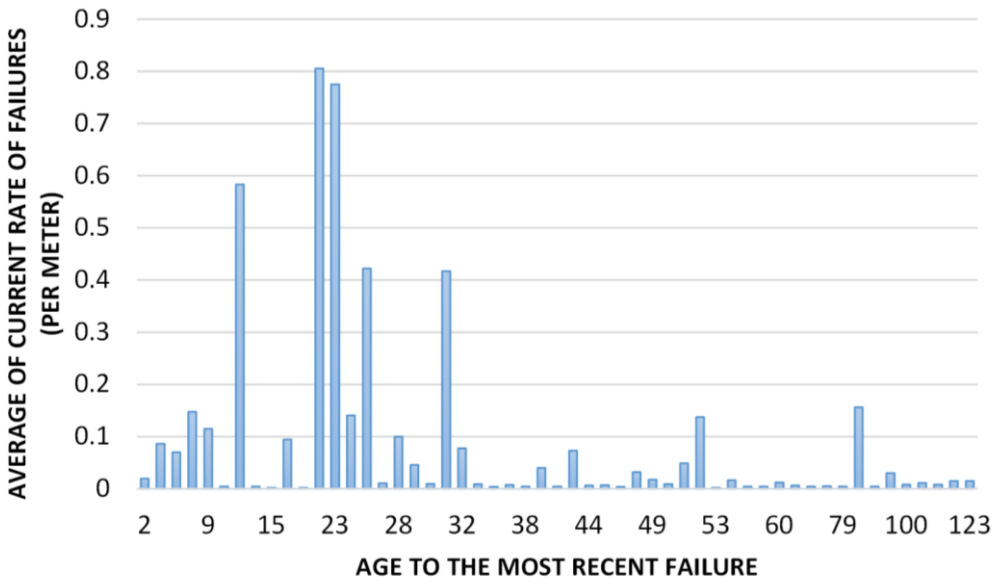


FIGURE 0.29 - AVERAGE OF CURRENT RATE OF FAILURE BASED ON AGE (REGION OF WATERLOO)

According to the results, random forest performed better than other algorithms. This can be seen with the R-Squared of 0.82 and the RMSE of 0.11. ElasticNet and ANN were the worst algorithms in this step, with the R-Squared of 0.07 and 0.17, respectively. XGBOOST, although not the best, had a comparatively good result compared to the random forest (TABLE 0.12).

TABLE 0.120.12 – REGRESSION METRICS – CURRENT RATE OF FAILURE (REGION OF WATERLOO)

Algorithm	RMSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron
ElasticNet	*0.26*	*N.A*	*0.07*	*N.A*
Random Forest	*0.11*	*N.A*	*0.82*	*N.A*

XGBOOST	*0.12*	*N.A*	*0.80*	*N.A*
ANN	*0.25*	*N.A*	*0.17*	*N.A*

* AM = All Materials

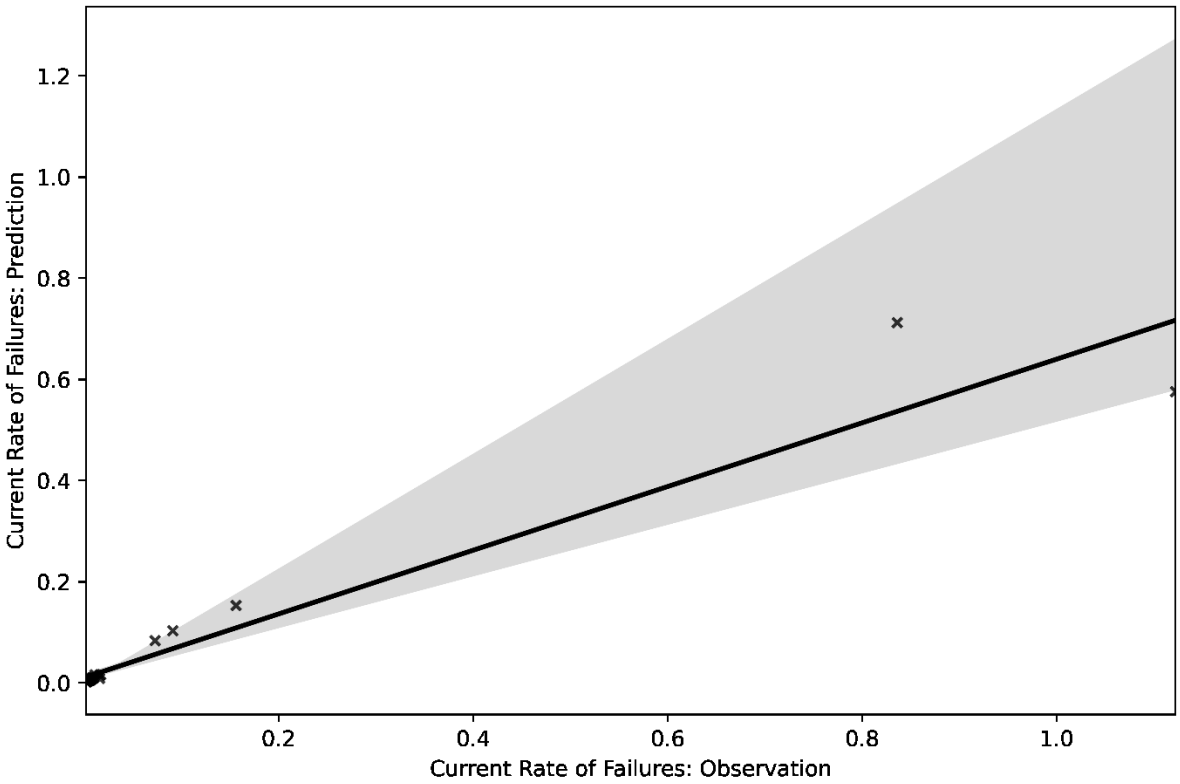


FIGURE 0.30 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – RANDOM FOREST) – REGION OF WATERLOO

Region of Durham

- ***Age at First Failure***

The final regression file for the Region of Durham includes 1,221 pipes, including different input variables; surface type, material, length, lining material, diameter, protection status, lining status, age, lining age, and protection age.

Two primary materials account for almost 95% of the entire network. The first one is cast iron pipe, making up 60.85%, and ductile iron that makes up 35.30% of the recorded failures. Other materials with insignificant failure can be noticed within the given pie chart (Figure 0.31).

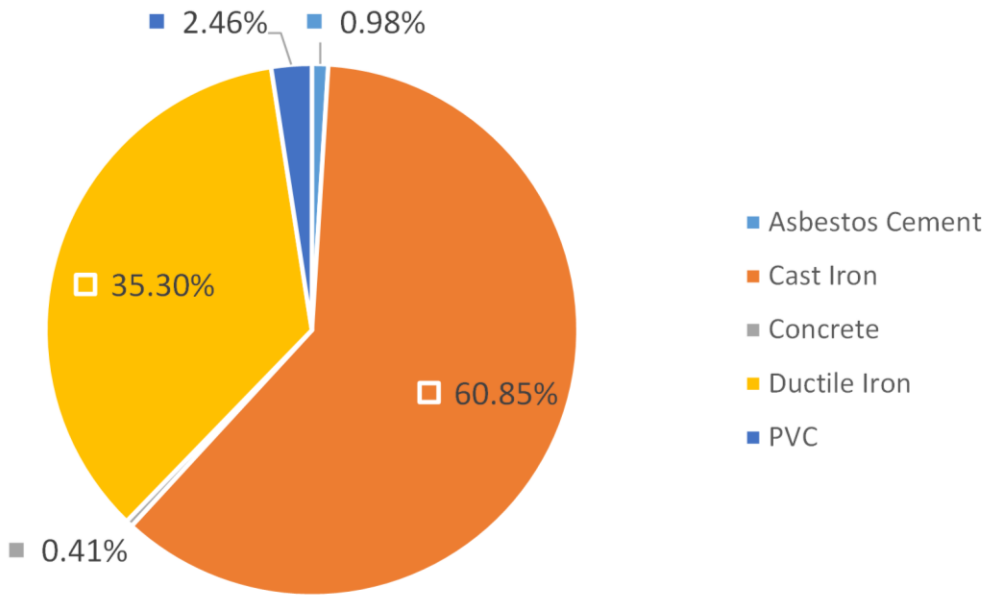


FIGURE 0.31 - PERCENTAGE OF EACH MATERIAL WITHIN REGRESSION ANALYSIS (AGE TO FIRST FAILURE) – REGION OF DURHAM

Cast iron with an average age of 38.48 experienced failures later in its life cycle compared to other materials. Concrete, in this case, has the youngest average age, which is 14.2. Asbestos cement, ductile iron, and PVC follow cast iron with an average of 36.17, 24.47, and 16.97 years. A box plot is created to show how age to the first failure is distributed among different materials (Figure 0.32).

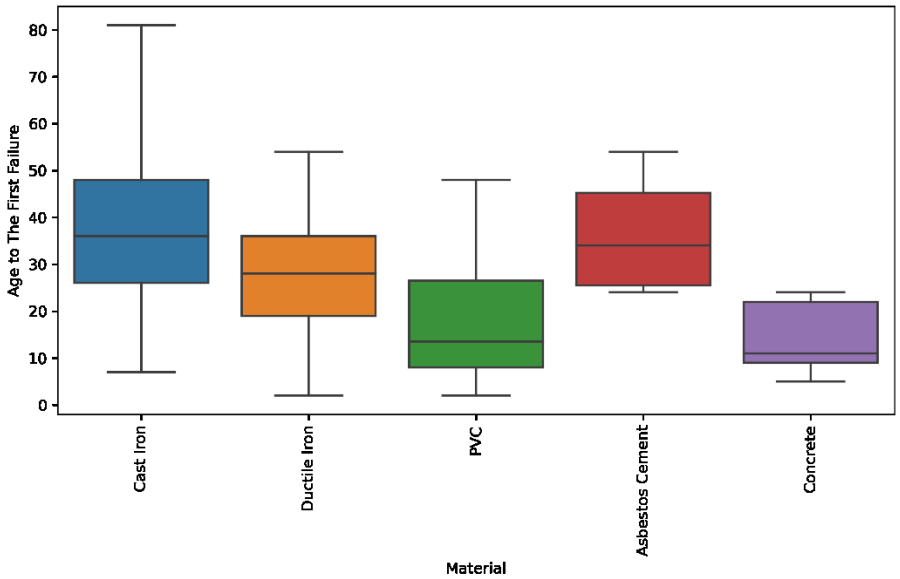


FIGURE 0.32- DISTRIBUTION OF AGE TO FIRST FAILURE BASED ON TYPE OF MATERIAL (REGION OF DURHAM)

For the Region of Durham, regression models did not indicate desirable results, as can be seen from the given table. Among these models, random forest with the RMSE of 14.9 and the R-Squared of 0.20 had the best performance for AM category (TABLE 0.13). This model also for cast iron pipe represented the best performance. Moreover, ElasticNet and ANN with an RMSE of 15.1 had better performance compared to XGBOOST. Having an R-Squared of 0.07 and an RMSE of 17.5, XGBOOST showed the worst performance for cast iron pipes. It should be noted that the accuracy of models declined after partitioning the dataset. A regression plot was created based on the random forest result comparing actual and predicted values (Figure 0.33).

TABLE 0.130.13 - REGRESSION METRICS (REGION OF DURHAM)

Algorithm	RMSE		MSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron	AM	Cast Iron
ElasticNet	*15.1*	*16.9*	*227.1*	*288.1*	*0.18*	*0.13*
Random Forest	*14.9*	*16.7*	*222.4*	*280.4*	*0.20*	*0.15*
XGBOOST	*15.2*	*17.5*	*230.3*	*306.9*	*0.17*	*0.07*
ANN	*15.1*	*17*	*228.3*	*288.5*	*0.18*	*0.13*

* AM = All Materials

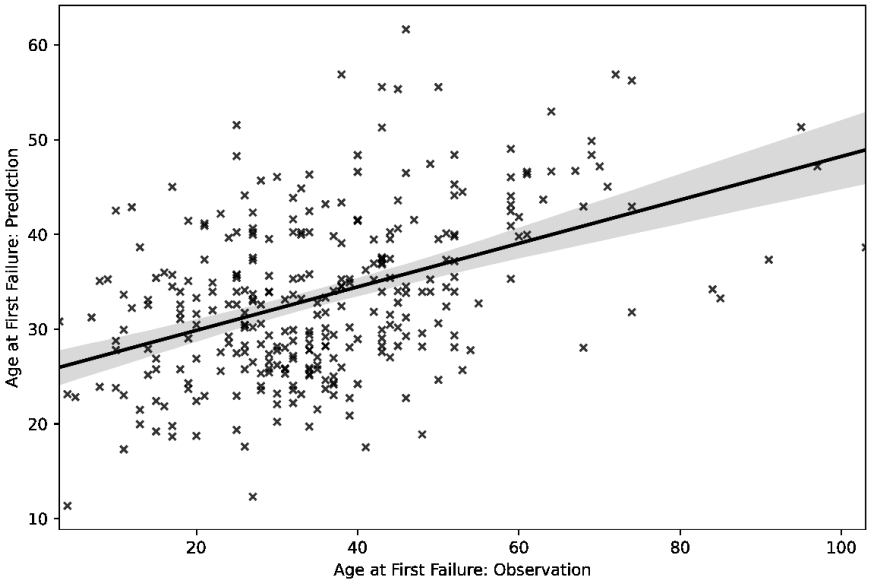


FIGURE 0.33 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – RANDOM FOREST) – REGION OF DURHAM

- ***Current Rate of Failure***

For the prediction of the current rate of failures, many input variables have been used for this network. These attributes consist of surface type, material, length, lining material, diameter, protection status, lining status, age, lining age, protection age, the previous rate of failures, and the current rate of failures.

The given chart indicates the average current rate of failure based on pipe age. Two peaks can be noticed within the graph. The first is at age ten which is around 0.06 per meter, and the other is around 86, which is relatively high, 0.14 per meter. This result indicates that the failure rate in the early and wear-out phases is higher than in the in-usage stage (Figure 0.34).

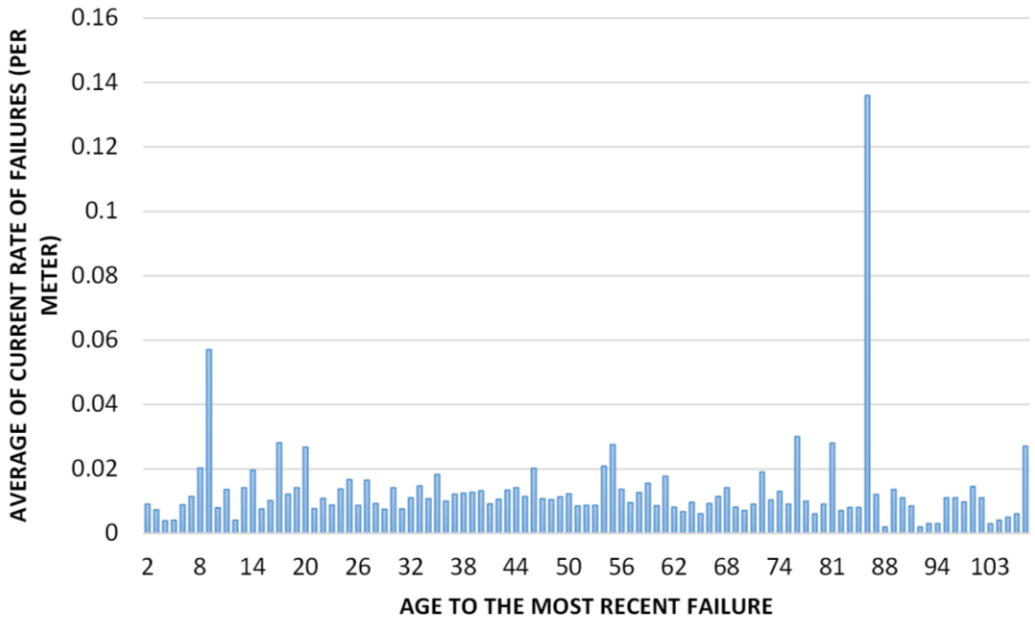


FIGURE 0.34 - AVERAGE OF CURRENT RATE OF FAILURE BASED ON AGE (REGION OF DURHAM)

The results for this step are relatively satisfactory, although the performance was not desirable for some models. For instance, the best models for the AM category were the random forest and XGBOOST, with similar RMSE and R-Squared scores, 0.012 and 0.78, respectively. For cast iron also, these two algorithms performed better than ElasticNet and ANN. Nevertheless, random forest with an R-squared of 0.85 indicated the best results for the cast iron group (TABLE 0.14). Moreover, the given regression plot indicates how random forest can fit the dataset in terms of prediction (Figure 0.35).

TABLE 0.140.14 – REGRESSION METRICS – CURRENT RATE OF FAILURE (REGION OF DURHAM)

Algorithm	RMSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron
ElasticNet	*0.024*	*0.012*	*0.09*	*0.18*

Random Forest	*0.012*	*0.005*	*0.78*	*0.85*
XGBOOST	*0.012*	*0.007*	*0.78*	*0.69*
ANN	*0.033*	*0.014*	*0.15*	*N.A*

* AM = All Materials

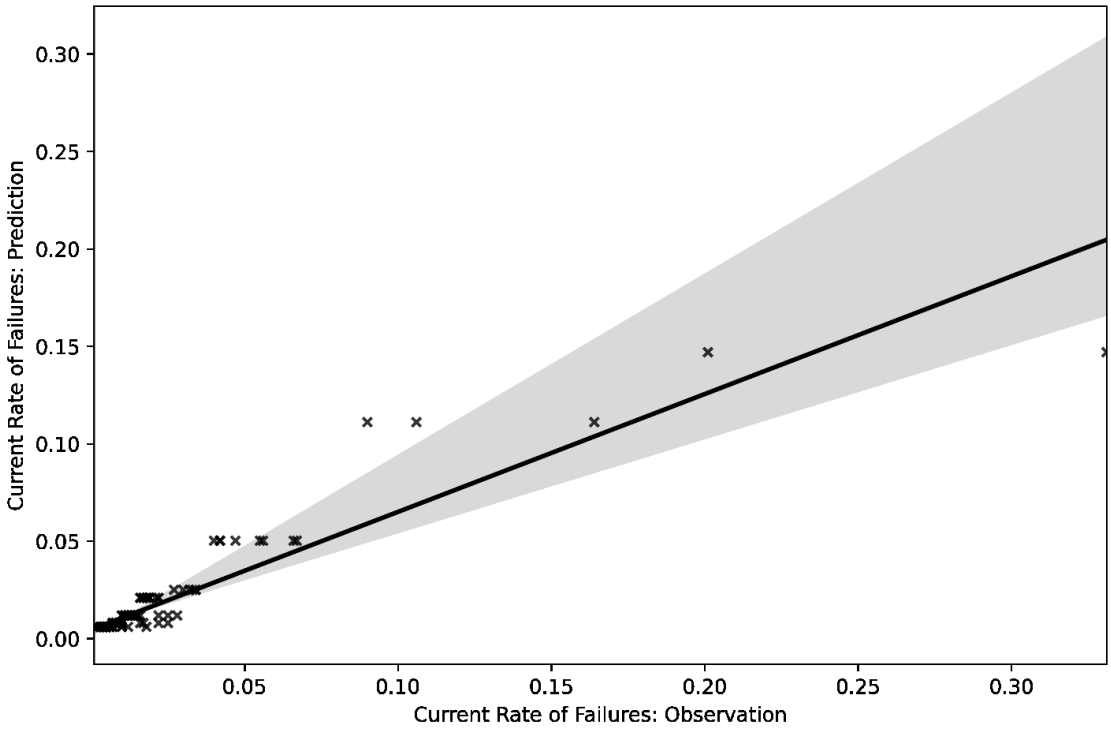


FIGURE 0.35 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – RANDOM FOREST) – REGION OF DURHAM

Calgary

- ***Age at First Failure***

Calgary, as discussed previously, owns one of the most robust datasets. This network includes 3,913 uniques pipes with at least one failure. This network`s dataset includes different input variables: material, length, diameter, age, coating status, protection status, anode type, average soil resistivity, and dead-end. It should be noted that a significant proportion of the file is allocated to the cast iron pipes, with a 66.39% contribution. Ductile iron accounts for almost 29.26% of the whole records, and asbestos cement with 2.68% is third. The given pie chart indicates the percentage of each material within the regression file (Figure 0.36).

Additionally, according to the available information, copper has the highest average age for the first failure, around 48.5 years. Meanwhile, concrete and cast iron are the following materials

with an average of 32.83 and 31.65, respectively. Finally, polyethylene has the lowest number, 8.5, indicating that this material is significantly prone to failure when in the early stage of the bathtub curve. The distribution of age to the first failure for all materials can be seen within the given box plot (Figure 0.37).

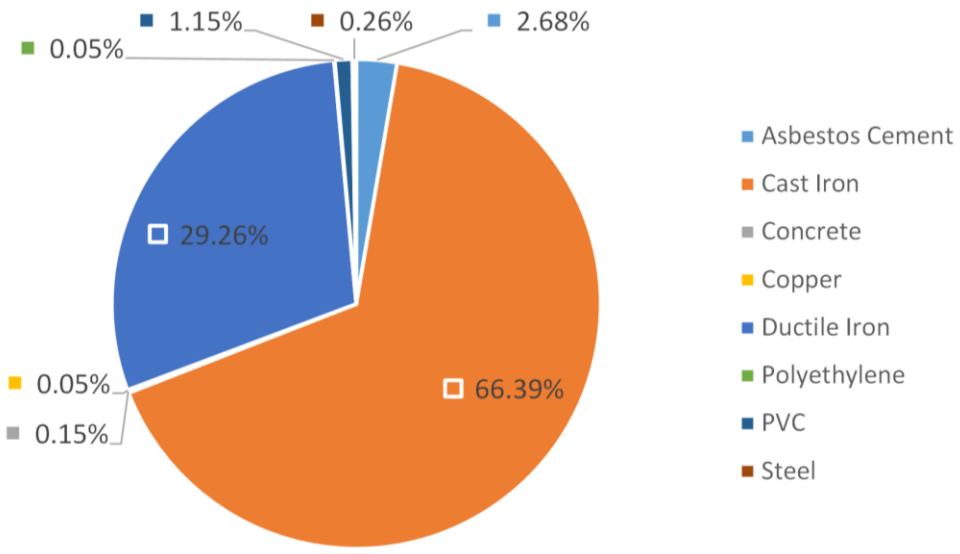


FIGURE 0.36 - PERCENTAGE OF EACH MATERIAL WITHIN REGRESSION ANALYSIS (AGE TO FIRST FAILURE) – CALGARY

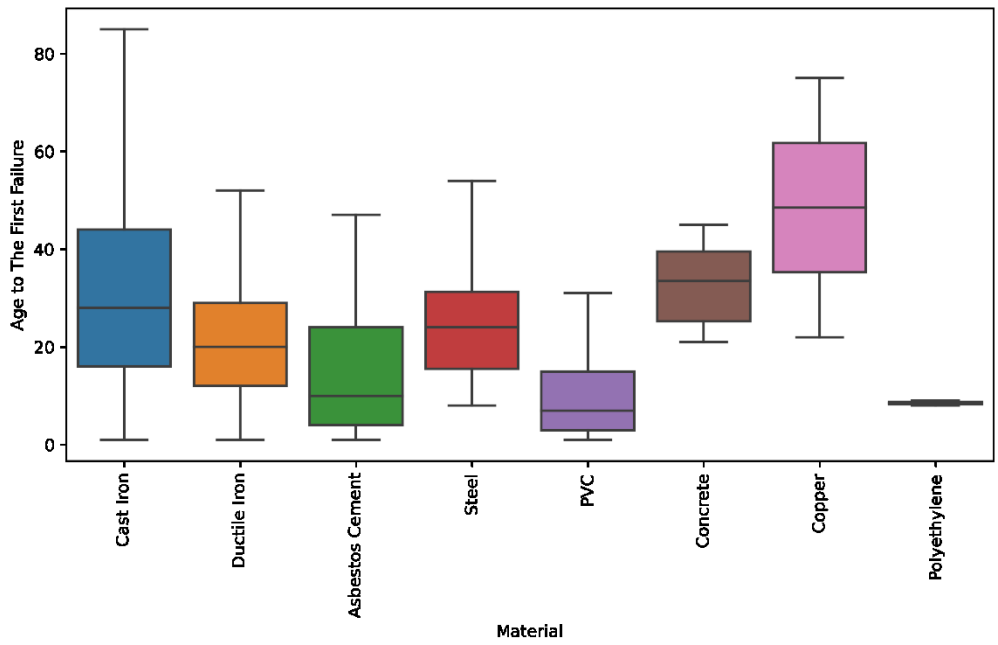


FIGURE 0.37 - DISTRIBUTION OF AGE TO FIRST FAILURE BASED ON TYPE OF MATERIAL (CALGARY)

Results for the Calgary network are provided in the given table. As can be seen, no models indicated a good performance, and they all performed somewhat similarly. For instance, the RMSE score for all models is around 16.5. Models for this city require further enhancement in order to become reliable and feasible in the real world (TABLE 0.15). As an example, the ElasticNet regression's low accuracy can be noticed in the given regression plot (Figure 0.38).

TABLE 0.150.15 - REGRESSION METRICS (CALGARY)

Algorithm	RMSE		MSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron	AM	Cast Iron
ElasticNet	*16.5*	*19.7*	*273.2*	*386.0*	*0.12*	*N.A*
Random Forest	*16.5*	*19.8*	*273.4*	*390.4*	*0.12*	*N.A*
XGBOOST	*16.7*	*20.2*	*277.5*	*406.1*	*0.11*	*N.A*
ANN	*16.6*	*19.8*	*274.7*	*390.1*	*0.12*	*N.A*

* AM = All Materials

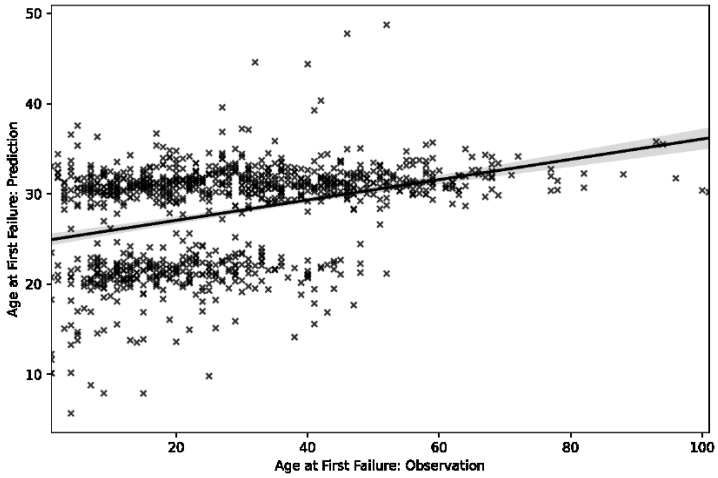


FIGURE 0.38 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – ELASTICNET) – CALGARY

- ***Current Rate of Failure***

For predicting the current rate of failure following attributes have been utilized: material, length, diameter, age, coating status, protection status, anode type, average soil resistivity, dead-end, the previous rate of failure, and the current rate of failure. As can be seen within the given bar chart, the younger the age to the most recent failure is, the higher the current rate of failure is in this network, experiencing a peak around 17. Nonetheless, there are a few peaks in the wear-out phase of the network. These peaks are for pipes that are around the age of 77 and over the age of 100. This trend could be related to the deterioration process of water mains during older age.

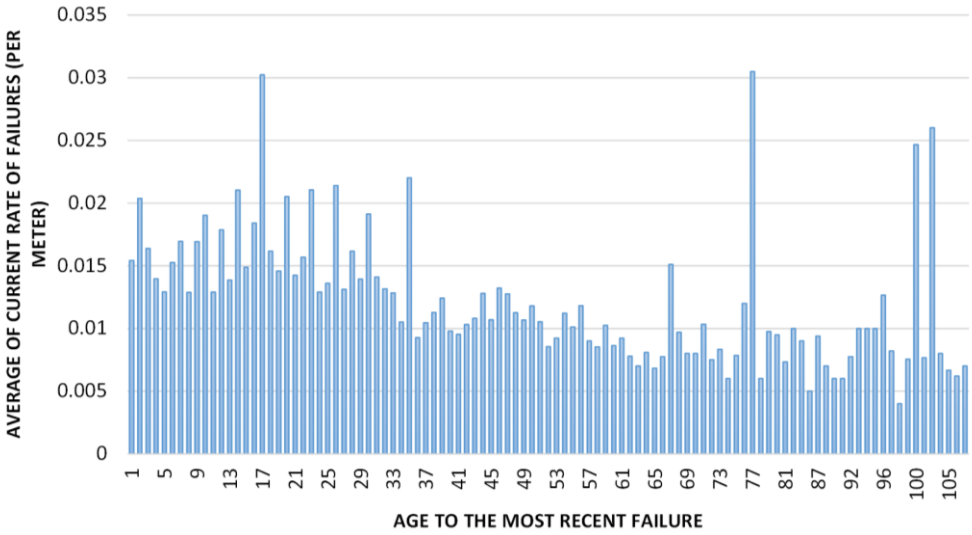


FIGURE 0.39 - AVERAGE OF CURRENT RATE OF FAILURE BASED ON AGE (CALGARY)

Comparing the results of regression models reveals high accuracy for tree-based algorithms; random forest, and XGBOOST. For example, for AM category, the RMSE score is 0.007 and 0.005 for random forest and XGBOOST, respectively, and the corresponding R-Squared values are 0.90 and 0.96. Thus, XGBOOST has proven to be the best model for predicting this network's current rate of failures.

Moreover, for the cast iron group, XGBOOST performed better than other models with an R-Squared of 0.98. On the other hand, elasticNet regression for both categories indicated the weakest accuracy. The given table below compares the regression metrics for all algorithms (TABLE 0.16).

TABLE 0.160.16 – REGRESSION METRICS – CURRENT RATE OF FAILURE (CALGARY)

Algorithm	RMSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron
ElasticNet	*0.02*	*0.016*	*0.15*	*0.21*
Random Forest	*0.007*	*0.004*	*0.90*	*0.94*
XGBOOST	*0.005*	*0.002*	*0.96*	*0.98*
ANN	*0.018*	*0.014*	*0.33*	*0.35*

* AM = All Materials

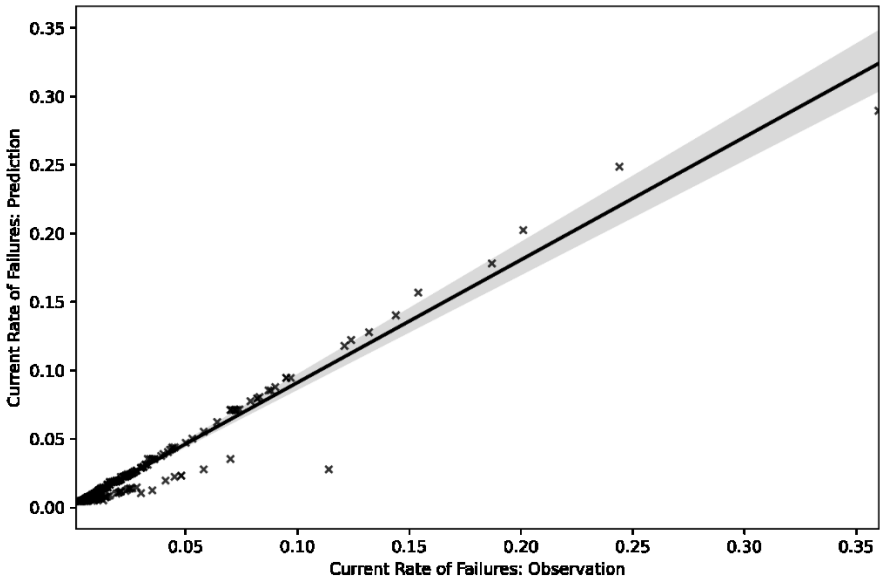


FIGURE 0.40 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – XGBOOST) – CALGARY

Vancouver

- ***Age at First Failure***

Although considered one of the largest networks across Canada, this network includes a small number of unique pipe IDs. Only 709 pipes recorded experienced at least one failure. The available attributes for this network in this step of the study are pipe depth, service type, diameter, length, material, coating material, lining material, and age to the first failure. Cast iron is the most frequent type of material within the regression file, with an 88.15% contribution. The following material is ductile iron that accounts for only 7.48% of total records. Steel pipes, also with 3.39%, are the following frequent material within the network.

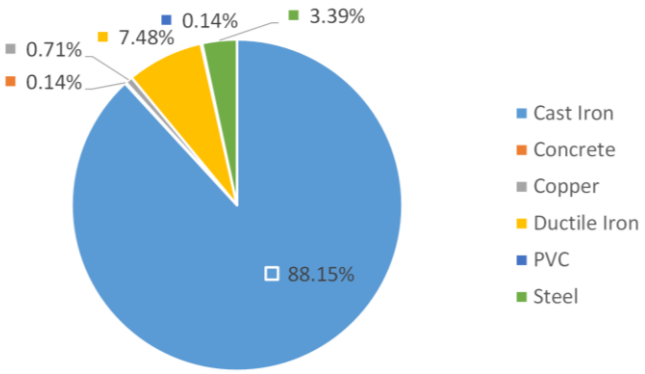


FIGURE 0.41 - PERCENTAGE OF EACH MATERIAL WITHIN REGRESSION ANALYSIS (AGE TO FIRST FAILURE) – VANCOUVER

Steel pipes have the highest average age to the first failure within the available information, with the value of 77.13, which is relatively high compared to other cities. Cast iron, also with 64.14, has a relatively high average compared to other materials within the network. The average age to the first failure for this network is somewhat high compared to other utilities (Figure 0.42).

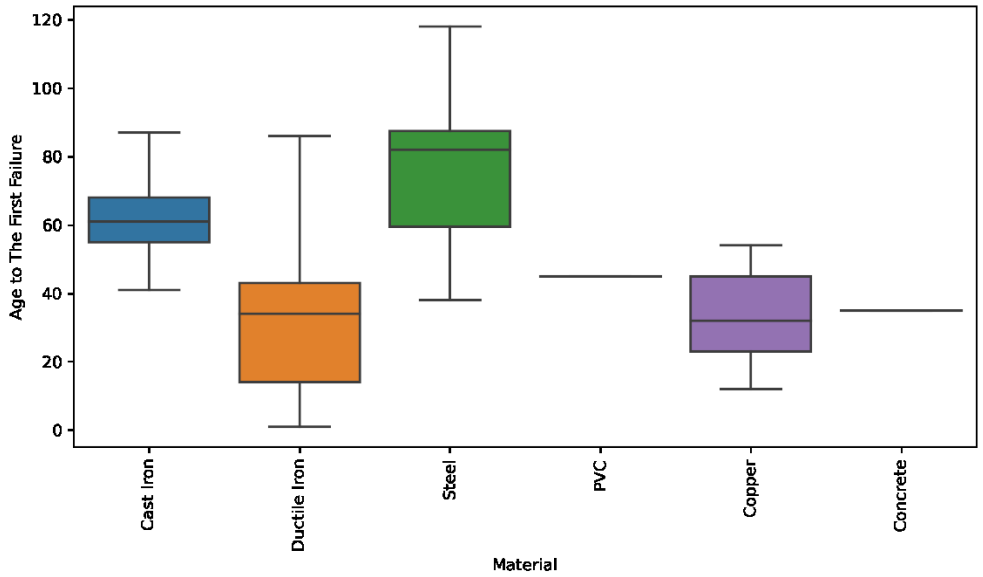


FIGURE 0.42 - DISTRIBUTION OF AGE TO FIRST FAILURE BASED ON TYPE OF MATERIAL (VANCOUVER)

According to the results for Vancouver, no regression algorithms were performed satisfactorily. On the other hand, ANN has the highest accuracy for AM category with the R-Squared of 0.39 and RMSE of 15.4. However, random forest indicated a better result for the cast iron group than other models with RMSE of 12.9 and R-Squared of 0.23. Nonetheless, it should be noted that these values are not as much reliable as they should be for applying to real-world cases. The given table below compares the results of these algorithms (TABLE 0.17). Looking at the given regression plot also emphasized the weakness of ANN, even as the best model for Calgary (Figure 0.43).

TABLE 0.170.17 - REGRESSION METRICS (VANCOUVER)

Algorithm	RMSE		MSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron	AM	Cast Iron
ElasticNet	*16.2*	*13.0*	*261.9*	*168*	*0.32*	*0.22*
Random Forest	*15.7*	*12.9*	*245.4*	*166.4*	*0.36*	*0.23*
XGBOOST	*16.4*	*13.6*	*268.6*	*185.7*	*0.30*	*0.14*
ANN	*15.4*	*13.1*	*236.6*	*172.3*	*0.39*	*0.20*

* AM = All Materials

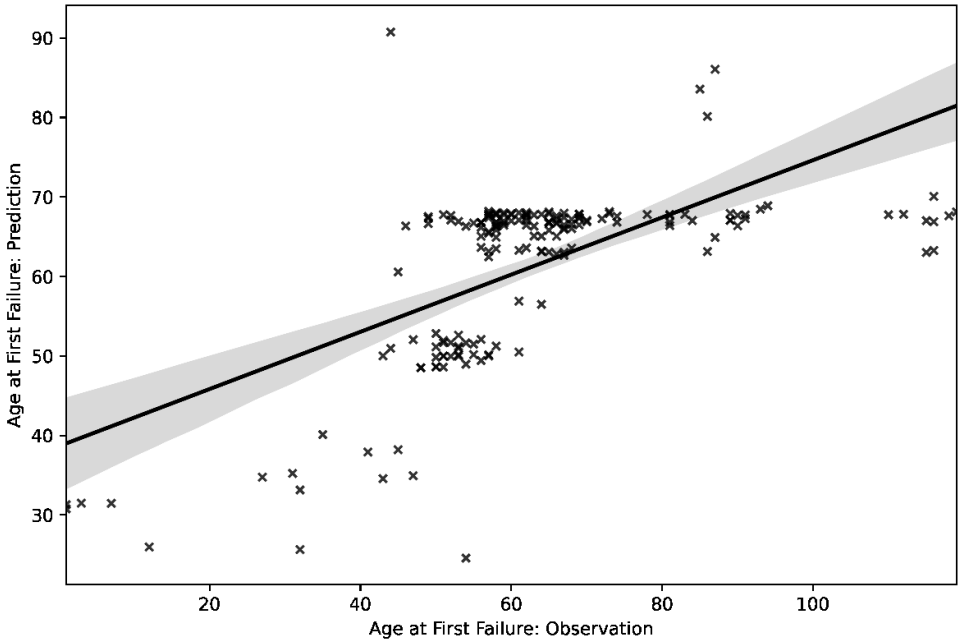


FIGURE 0.43 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – ANN) – VANCOUVER

- ***Current Rate of Failure***

The input variables for the current rate of failure in Vancouver are different from other cities since they do not include previous failure rates. This is because the first failure for each pipe is just reported in a specific year, which means that the most recent failure in this file is the first failure for all pipes. Therefore, only the current rate of failure has been calculated and added to the dataset. Accordingly, the following are the input variables for the city of Vancouver: pipe depth, service type, diameter, length, material, coating material, lining material, the current rate of failures, and age at most recent failure. The Given bar chart indicates that younger pipes experienced more failure than older pipes (Figure 0.44).

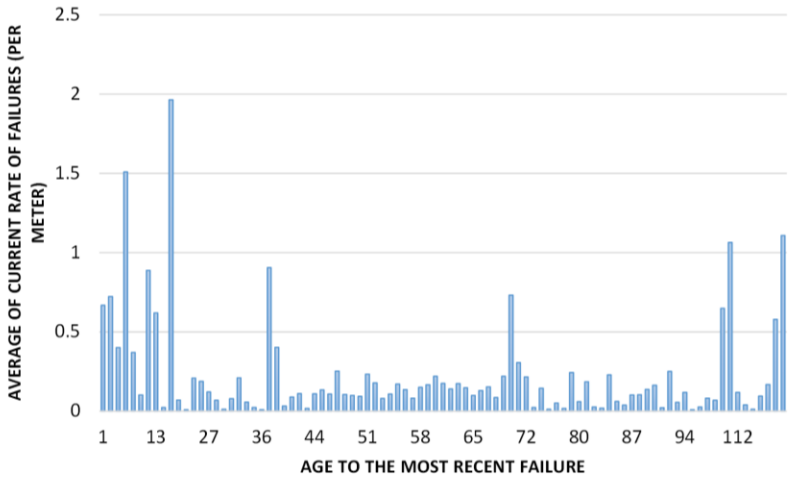


FIGURE 0.44 - AVERAGE OF CURRENT RATE OF FAILURE BASED ON AGE (VANCOUVER)

TABLE 0.180.18 – REGRESSION METRICS – CURRENT RATE OF FAILURE (VANCOUVER)

Algorithm	RMSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron
ElasticNet	*0.245*	*0.223*	*0.01*	*0.10*
Random Forest	*0.034*	*0.024*	*0.98*	*0.99*
XGBOOST	*0.036*	*0.027*	*0.98*	*0.97*
ANN	*0.090*	*0.079*	*0.88*	*0.89*

* AM = All Materials

Considering the results, however, regression models performed better than the first step of the study, which was the prediction of the age at the first failure. For instance, random forest and XGBOOST showed a significantly high performance with an R-Squared of 0.98. This metric for the ANN model is 0.88, which is relatively desirable. It should be noted that ElasticNet indicated a low accuracy for AM category, with an R-Squared of 0.01.

On the other hand, random forest resulted in better accuracy for the cast iron group with an R-Squared of 0.99, which is significantly high. However, no considerable improvement can be seen with partitioning the dataset based on material in the cast iron group. Random forest and ANN`s R-Squared increased by merely 0.01. The given figure shows the goodness of fit for the random forest model (Figure 0.45).

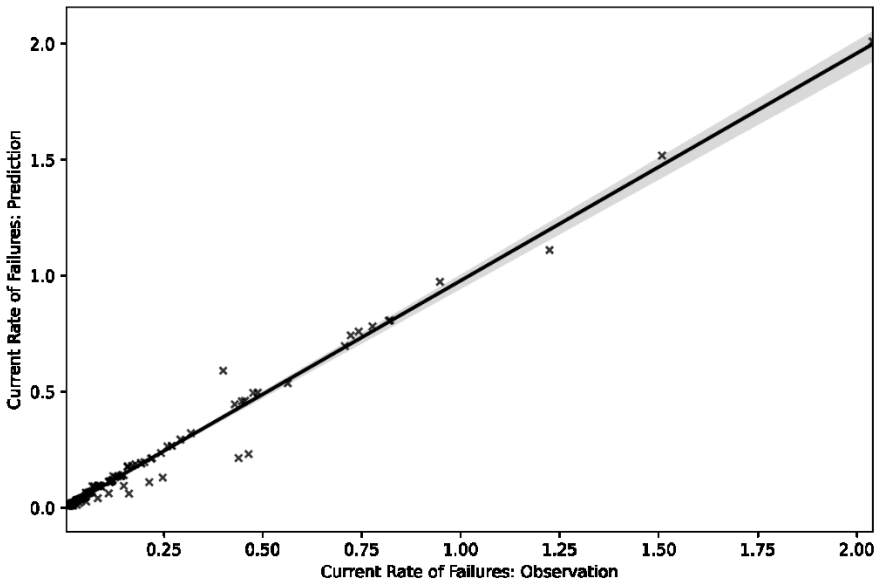


FIGURE 0.45 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – RANDOM FOREST) – VANCOUVER

Victoria

- ***Age at First Failure***

Victoria water network includes 397 unique pipes with at least one failure record. These pipes are provided with several attributes: material, HGL, diameter, length, lining material, lining status, and age to the first failure. For this part of the study, cast iron accounts for almost 75% of total failure, followed by ductile iron with an 18.89% contribution. As can be seen in the given pie chart, other materials make up for a small proportion of recorded failures (Figure 0.46).

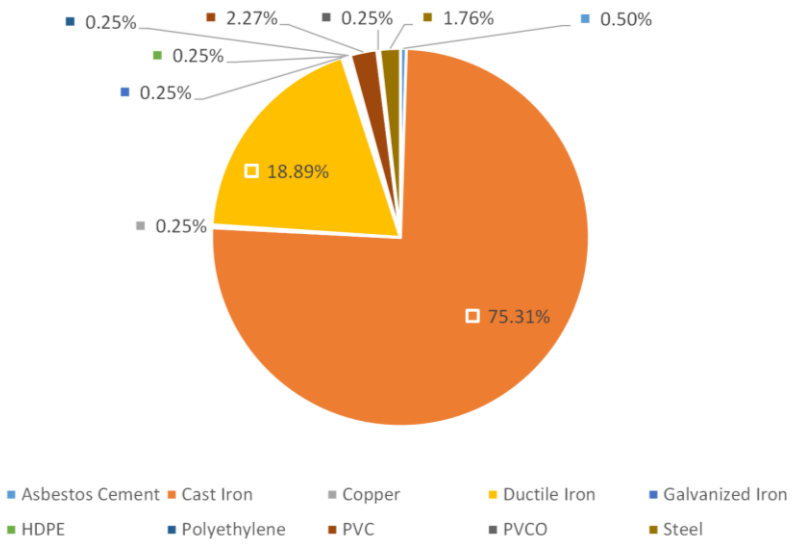


FIGURE 0.46 - PERCENTAGE OF EACH MATERIAL WITHIN REGRESSION ANALYSIS (AGE TO FIRST FAILURE) – VICTORIA

The given box plot provides information about age distribution to the first failure based on different materials (Figure 0.47). For example, steel pipes recorded the highest value, which is around 84 years. This value for cast iron and ductile iron is 57 and 32, respectively. The average failure age for other materials can be found in the given chart.

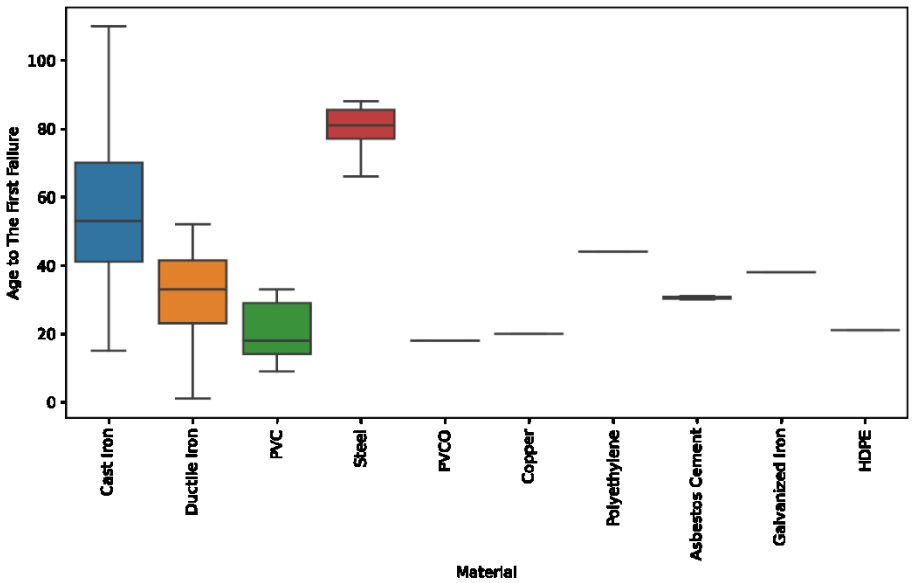


FIGURE 0.47 - DISTRIBUTION OF AGE TO FIRST FAILURE BASED ON TYPE OF MATERIAL (VICTORIA)

Results from regression analysis revealed that random forest, for AM category, had the best performance with an RMSE of 17.3 and R-Squared of 0.38. XGBOOST and ANN with R-Squared

of 0.33 and 0.38, respectively, are in the following positions. For cast-iron pipes, however, ElasticNet regression provided a better result with an R-Squared of 0.30. ANN, for cast iron, did not have a desirable result (TABLE 0.19).

TABLE 0.190.19 - REGRESSION METRICS (VICTORIA)

Algorithm	RMSE		MSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron	AM	Cast Iron
ElasticNet	*18.4*	*19.2*	*339.5*	*369.5*	*0.30*	*0.30*
Random Forest	*17.3*	*20.0*	*298.1*	*401.0*	*0.38*	*0.24*
XGBOOST	*17.9*	*22.7*	*322.6*	*517.7*	*0.33*	*0.02*
ANN	*18.6*	*23.2*	*346.1*	*552.9*	*0.28*	*0.00*

* AM = All Materials

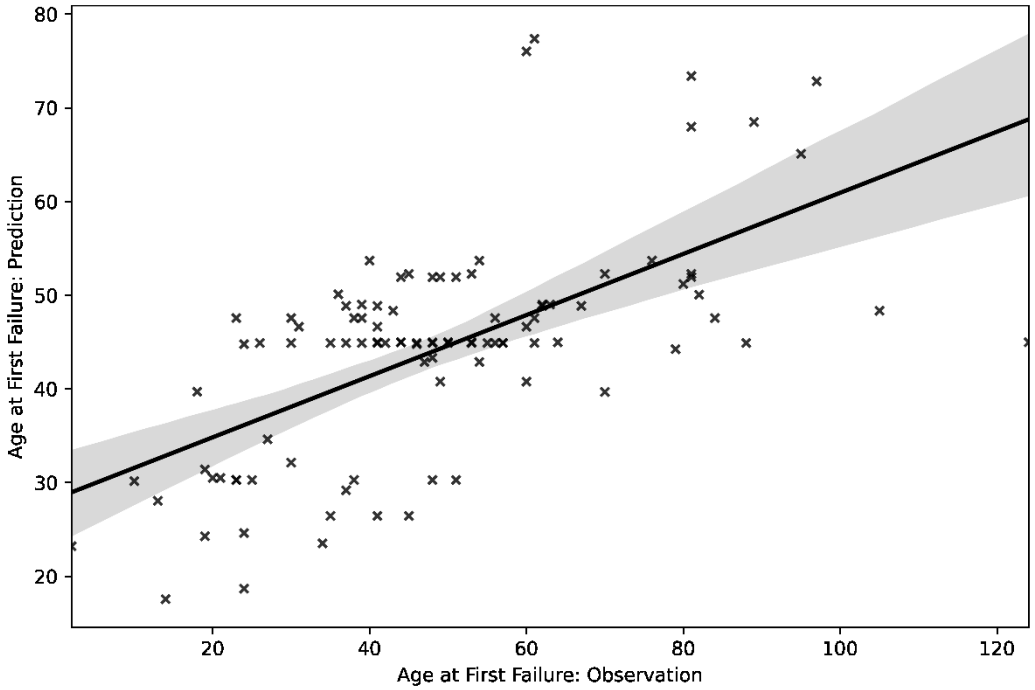


FIGURE 0.48 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – RANDOM FOREST) – VICTORIA

- ***Current Rate of Failure***

The current rate of failure was also analyzed for the city of Victoria. The given histogram shows the average current rate of failure based on the age at most recent failure (Figure 0.49). As shown, a fluctuation can be noticed for different ages. Seemingly, the rate of failure during the early stage and wear-out stage of the bathtub curve is higher than the in-usage period. A range of input variables has been used for this part of the analysis in this network: material, HGL,

diameter, length, lining material, lining status, age, the previous rate of failure, and the current rate of failure (dependent variable).

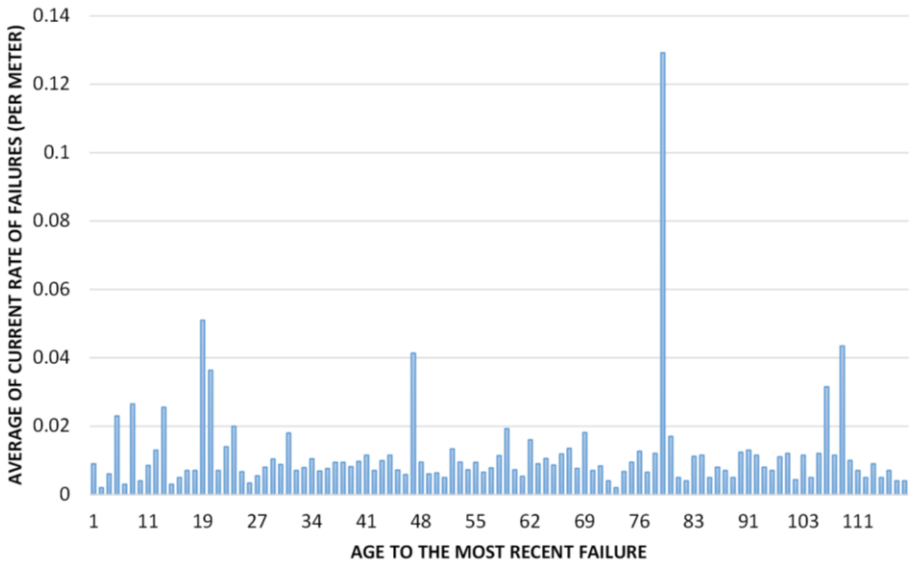


FIGURE 0.49 - AVERAGE OF CURRENT RATE OF FAILURE BASED ON AGE (VICTORIA)

ElasticNet regression did not present a desirable performance for this network, and with an R-Squared of 0.02, it was the weakest model. Tree-based models, on the other hand, indicated higher scores. XGBOOST, with an RMSE of 0.003 and an R-Squared of 0.90, was able to predict the rate of failures more accurately. The random forest also with 0.84 R-Squared, after XGBOOST, was the best predictive model in Victoria. Additionally, for cast iron pipe, XGBOOST and random forest with an R-Sqaured of 0.35 and 0.27, respectively, indicated a better performance than other algorithms. Interestingly, homogeneity decreased the power of the model in prediction. Provided is the table that shows the regression metrics for all models (TABLE 0.20). Figure 0.50 also plots the predicted current rate of failure versus the actual current rate of failure.

TABLE 0.200.20 – REGRESSION METRICS – CURRENT RATE OF FAILURE (VICTORIA)

Algorithm	RMSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron
ElasticNet	*0.01*	*0.05*	*0.02*	*0.03*
Random Forest	*0.004*	*0.04*	*0.84*	*0.27*
XGBOOST	*0.003*	*0.04*	*0.90*	*0.35*
ANN	*0.015*	*0.05*	*N.A*	*0.03*

* AM = All Materials

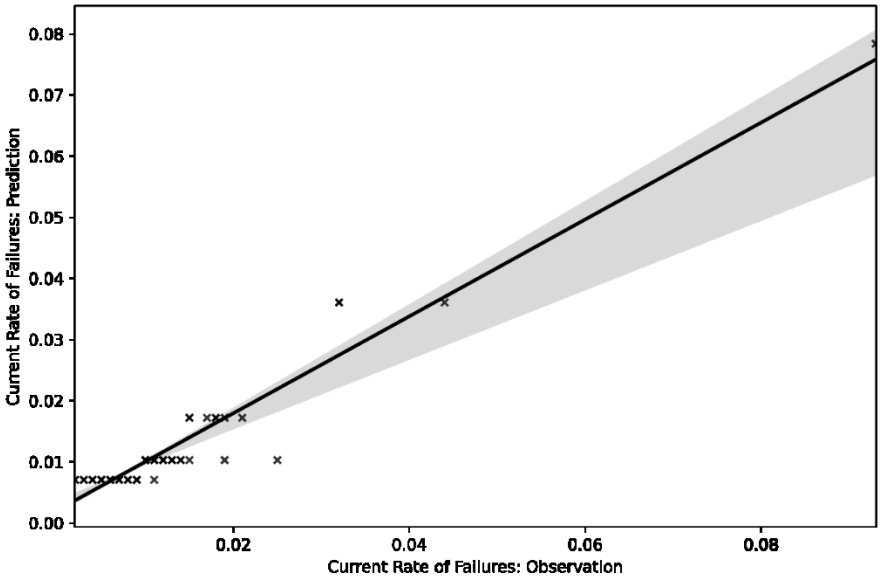


FIGURE 0.50 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – XGBOOST) – VICTORIA

Halifax

- ***Age at First Failure***

Halifax network consists of 1,835 unique pipes that experienced at least one failure. Length, diameter, material, age, lining status, and lining material were used as input variables for the regression analysis of Halifax. Based on the given information, 85.50% of the total pipes in the regression file are related to cast iron material. Ductile iron is the following material that failed more frequently than other materials, with 12.26% of failures recorded. Conversely, PVC with a 0.98% contribution is among the lowest recorded failures (Figure 0.51).

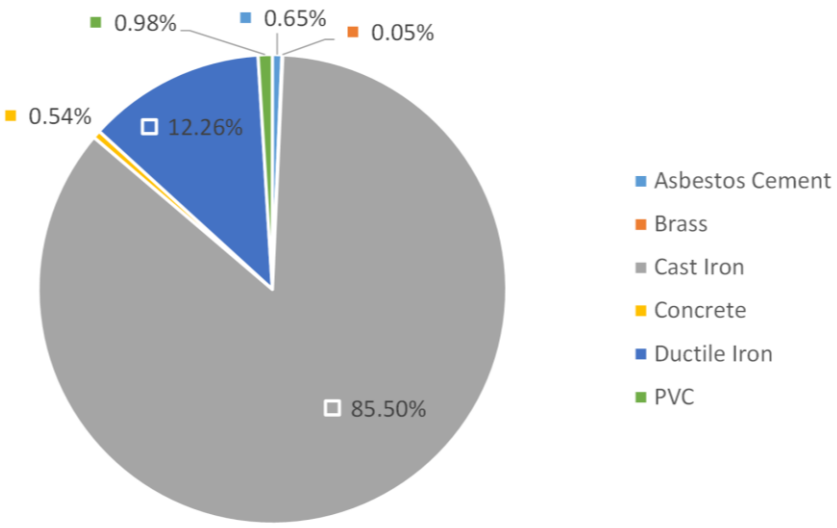


FIGURE 0.51 - PERCENTAGE OF EACH MATERIAL WITHIN REGRESSION ANALYSIS (AGE TO FIRST FAILURE) – HALIFAX

As previously mentioned, the number of years to the first failure is the target of the prediction in this step. Accordingly, the given box plot was created to show the distribution of age at the first failure for different materials. For example, cast iron and asbestos cement pipes have the highest average age at first failure, 42 and 37.5, respectively. On the other hand, PVC with 16.33 was the pipe with the lowest age at first failure. The distribution of age for other materials is provided as follows (Figure 0.52).

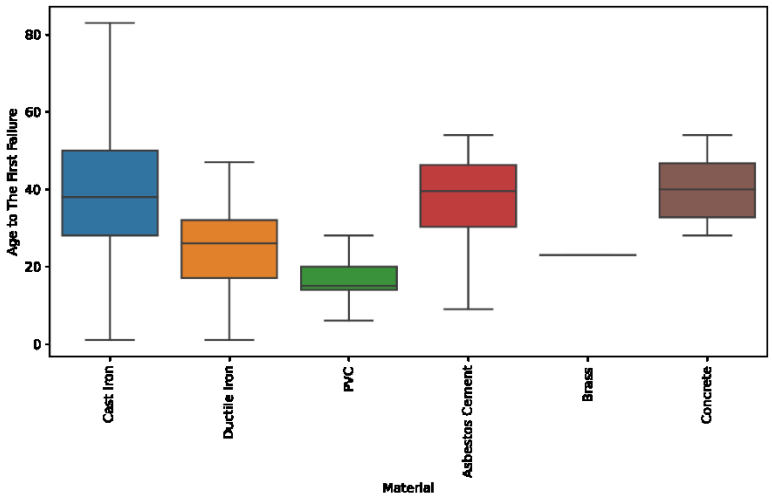


FIGURE 0.52 - DISTRIBUTION OF AGE TO FIRST FAILURE BASED ON TYPE OF MATERIAL (HALIFAX)

Regression models did not provide satisfactory results for the prediction of age at first failure. However, looking at the graph reveals that tree-based models performed better compared to ElasticNet and ANN models. For AM category, random forest with an RMSE of 17.2 and an R-

Squared of 0.31 was the best model. Conversely, ANN indicated the worst performance with the R-Squared of 0.08.

Furthermore, for the cast iron category, random forest with an RMSE and R-Squared of 18.5 and 0.16, respectively, provided better results than other models. The given table depicts more information about the results achieved in this part of the study (TABLE 0.21).

TABLE 0.210.21 - REGRESSION METRICS (HALIFAX)

Algorithm	RMSE		MSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron	AM	Cast Iron
ElasticNet	*18.5*	*19.2*	*340.6*	*367.266*	*0.20*	*0.09*
Random Forest	*17.2*	*18.5*	*294.2*	*341.2*	*0.31*	*0.16*
XGBOOST	*17.5*	*18.6*	*306.1*	*346.7*	*0.28*	*0.14*
ANN	*19.8*	*18.6*	*393.7*	*346.1*	*0.08*	*0.14*

* AM = All Materials

The given plot compares the predicted age at the first failure and the actual age at the first failure for the Halifax network (Figure 0.53).

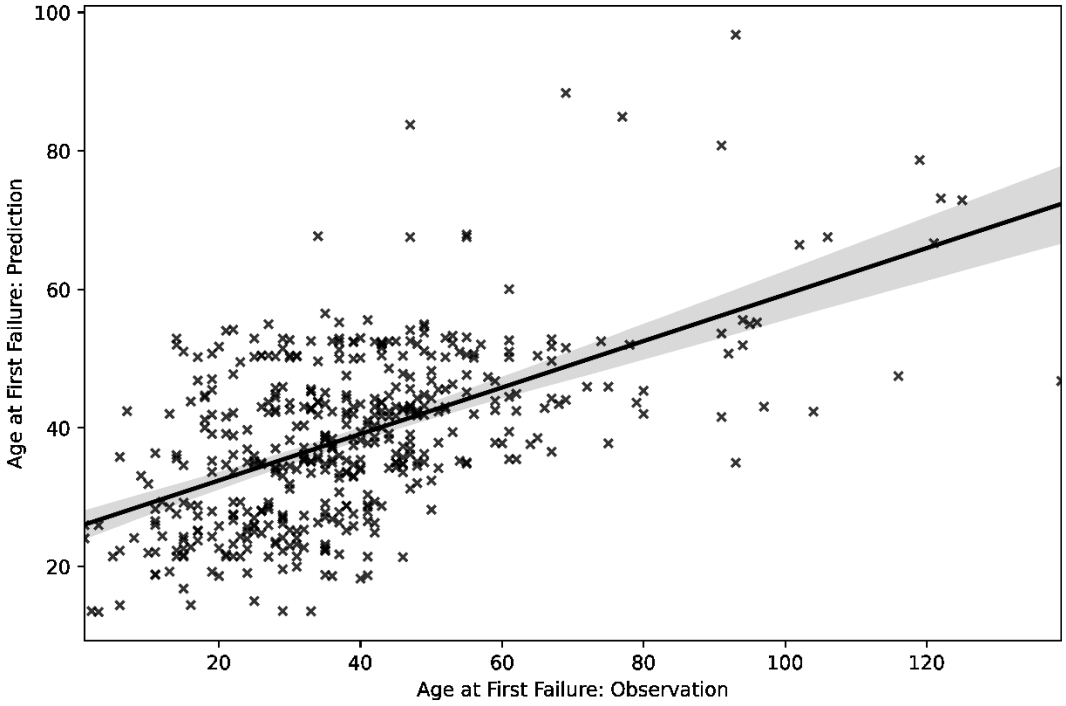


FIGURE 0.53 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – RANDOM FOREST) – HALIFAX

- ***Current Rate of Failure***

As shown in the given chart, the average current rate of failure is higher among the younger pipes (Figure 0.54). This can be seen with a peak around the age of 6 and 20. However, it should be mentioned that the current rate of failure is highly related to the length, and considering age alone is not as reliable. For this part of the study, length, diameter, material, lining material, lining status, the current rate of failure, the previous rate of failure, and age were employed as input variables.

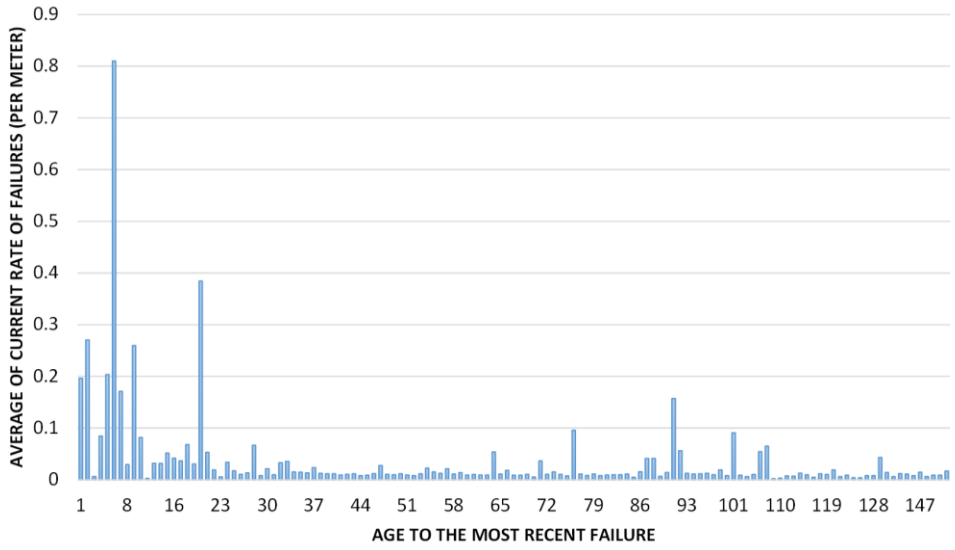


FIGURE 0.54 - AVERAGE OF CURRENT RATE OF FAILURE BASED ON AGE (HALIFAX)

For this network, random forest, XGBOOST, and ANN indicated satisfactory results. However, XGBOOST performed relatively better than the other two algorithms for both categories; AM and cast iron. The R-Squared and RMSE for XGBOOST are 0.92 and 0.029, respectively. The accuracy of XGBOOST declined when analyzing cast iron pipes, indicating that other criteria are also required for making uniform groups, and material alone is not enough. For instance, R-Squared decreased from 0.92 to 0.56 for XGBOOST model. Other results can be found in the given table (TABLE 0.22).

TABLE 0.220.22 – REGRESSION METRICS – CURRENT RATE OF FAILURE (HALIFAX)

Algorithm	RMSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron
ElasticNet	*0.10*	*0.11*	*0.08*	*0.01*
Random Forest	*0.043*	*0.091*	*0.83*	*0.34*
XGBOOST	*0.029*	*0.744*	*0.92*	*0.56*

ANN	*0.052*	*0.093*	*0.75*	*0.31*

* AM = All Materials

Figure 0.55 was prepared based on the random forest results. As can be seen, this model was able to predict the current rate of failure with relatively desirable accuracy.

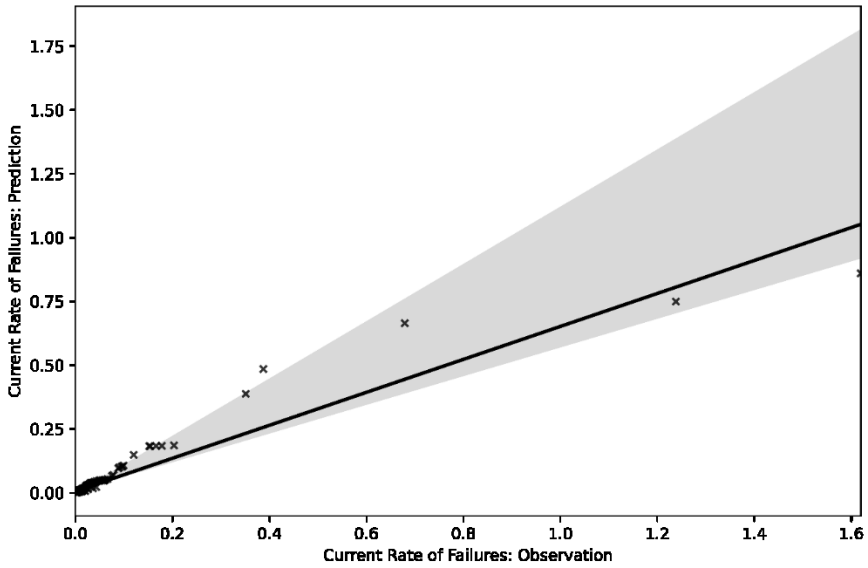


FIGURE 0.55 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – XGBOOST) – HALIFAX

St. John's

- ***Age at First Failure***

After cleaning and preparing the regression dataset, St. John's includes 833 unique pipes with at least one historical failure. There are different input variables along with this dataset, such as material, length, diameter, roughness, and age at the first failure.

Furthermore, cast iron pipes with 83.19% of recorded failures have the highest contribution. Ductile iron follows this material and accounts for 15.25% of total data points. The given pie chart provides more information about the frequency of all materials within the St. John's network (Figure 0.56).

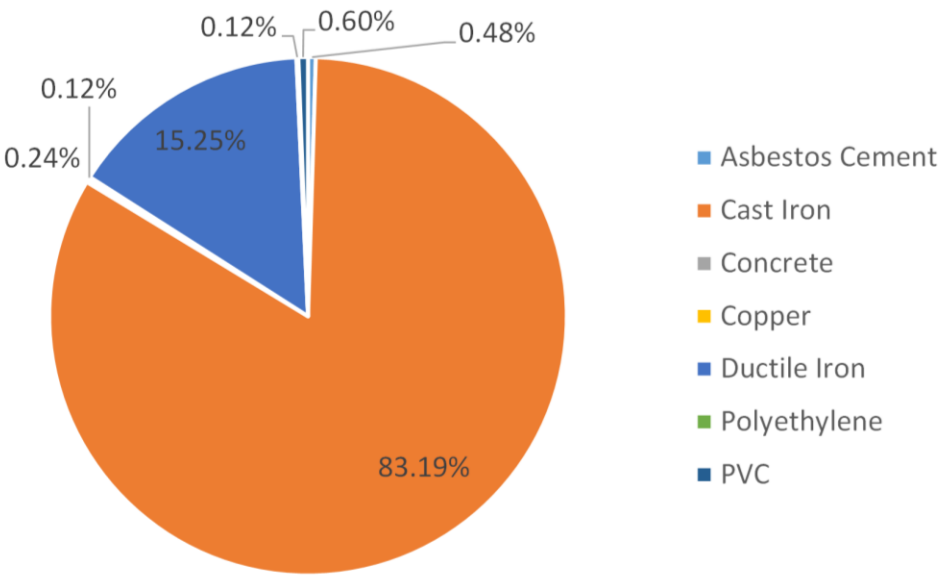


FIGURE 0.56 - PERCENTAGE OF EACH MATERIAL WITHIN REGRESSION ANALYSIS (AGE TO FIRST FAILURE) – ST.JOHN'S

Cast iron has the highest value considering the average age at first failure, which is 51.39 years. PVC pipes, however, experienced a minimum number of years to the first failure in this network. The average for PVC is around seven years. Moreover, ductile iron and asbestos cement pipes are after cast iron, with values of 25.25 and 23.5, respectively (Figure 0.57).

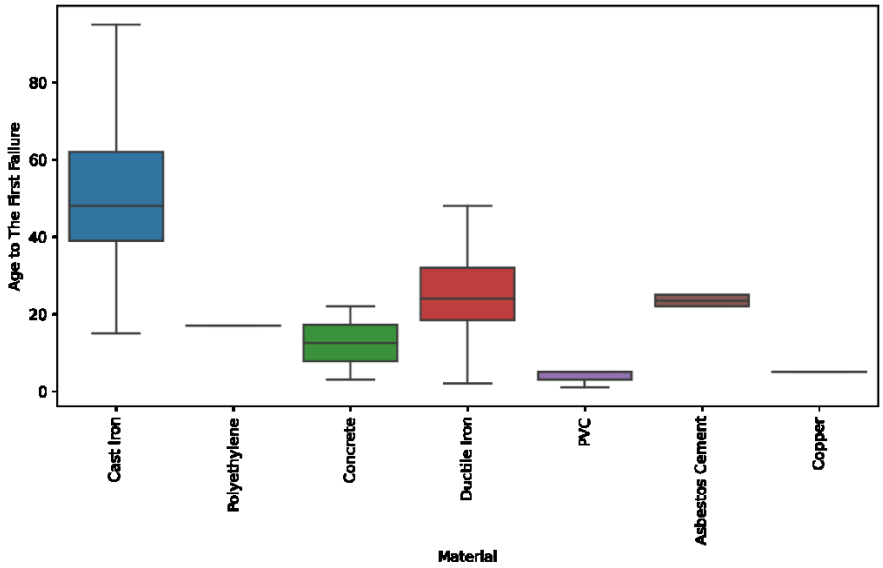


FIGURE 0.57 - DISTRIBUTION OF AGE TO FIRST FAILURE BASED ON TYPE OF MATERIAL (ST.JOHN'S)

The given table below compares the results of different regression models (TABLE 0.23). The ANN model performed better for AM category than others, as seen from RMSE and MSE scores,

12.3 and 352.5, respectively. However, this model did not provide a logical R-Squared score. On the other hand, the results for cast iron pipes were relatively close. Nonetheless, random forest with an RMSE score ok 18.5 and an R-Squared of 341.8 indicated a better accuracy than other models.

TABLE 0.230.23 - REGRESSION METRICS (ST. JOHN'S)

Algorithm	RMSE		MSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron	AM	Cast Iron
ElasticNet	*18.1*	*19.1*	*328.6*	*365.3*	*0.23*	*N.A*
Random Forest	*17.6*	*18.5*	*309.7*	*341.8*	*0.27*	*0.06*
XGBOOST	*18.0*	*20.4*	*323.3*	*417.7*	*0.24*	*N.A*
ANN	*12.3*	*18.8*	*151.1*	*352.5*	*N.A*	*0.035*

* AM = All Materials

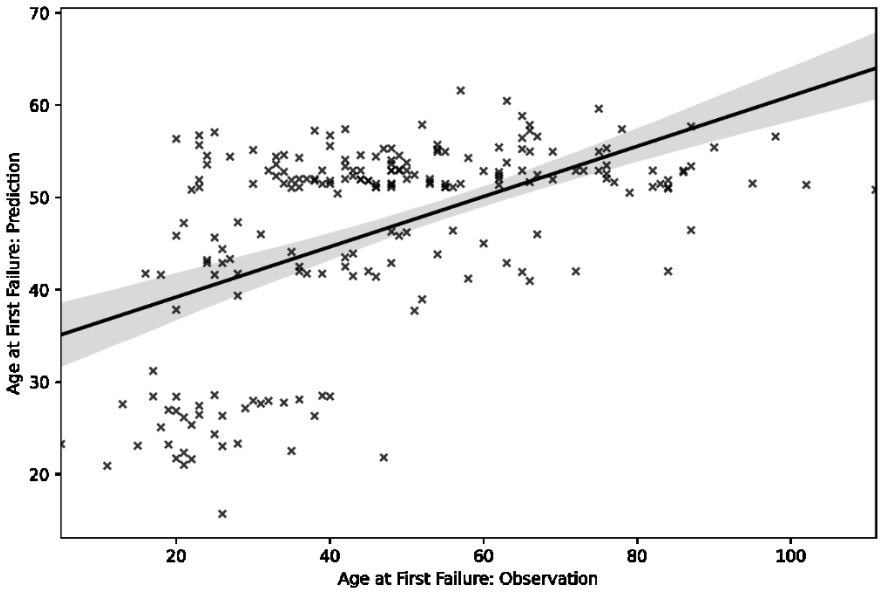


FIGURE 0.58 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – RANDOM FOREST) – ST.JOHN'S

- ***Current Rate of Failure***

Regression analysis indicated that the current rate of failures for young-age pipes and old-age pipes is higher compared to mid-age pipes. A peak can be seen for age around 75, with the average current rate of failure of 0.25 per meter. In this step, several input variables were employed for the prediction, such as diameter, roughness, length, material, age, the previous rate of failure, and the current failure rate. The given figure shows the average of the current rate of failure based on the age at most recent failure in St. John's (Figure 0.59)

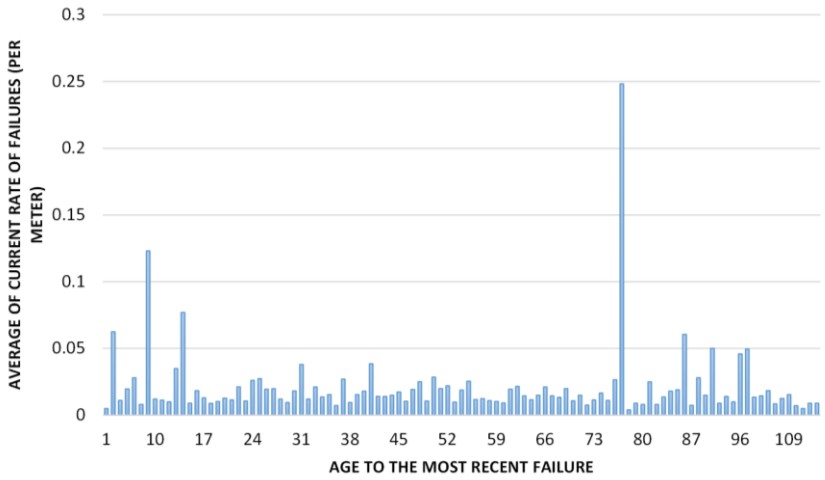


FIGURE 0.59 - AVERAGE OF CURRENT RATE OF FAILURE BASED ON AGE (ST. JOHN'S)

For AM category, XGBOOST indicated the best performance among all algorithms, with an RMSE of 0.006 and an R-squared of 0.95. For the cast iron group, also, this model performed better than others with an R-Squared of 0.98. Moreover, the performance of this model and random forest increased when used for the cast iron group. For XGBOOST from 0.95 to .098, and for random forest from 0.84 to 0.89. However, for the ANN model and Elasticnet, R-Squared worsened. Results of this part can be seen in the given table (TABLE 0.24).

TABLE 0.240.24 – REGRESSION METRICS – CURRENT RATE OF FAILURE (ST. JOHN'S)

Algorithm	RMSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron
ElasticNet	*0.025*	*0.024*	*0.25*	*0.20*
Random Forest	*0.011*	*0.009*	*0.84*	*0.89*
XGBOOST	*0.006*	*0.004*	*0.95*	*0.98*
ANN	*0.020*	*0.022*	*0.50*	*0.34*

* AM = All Materials

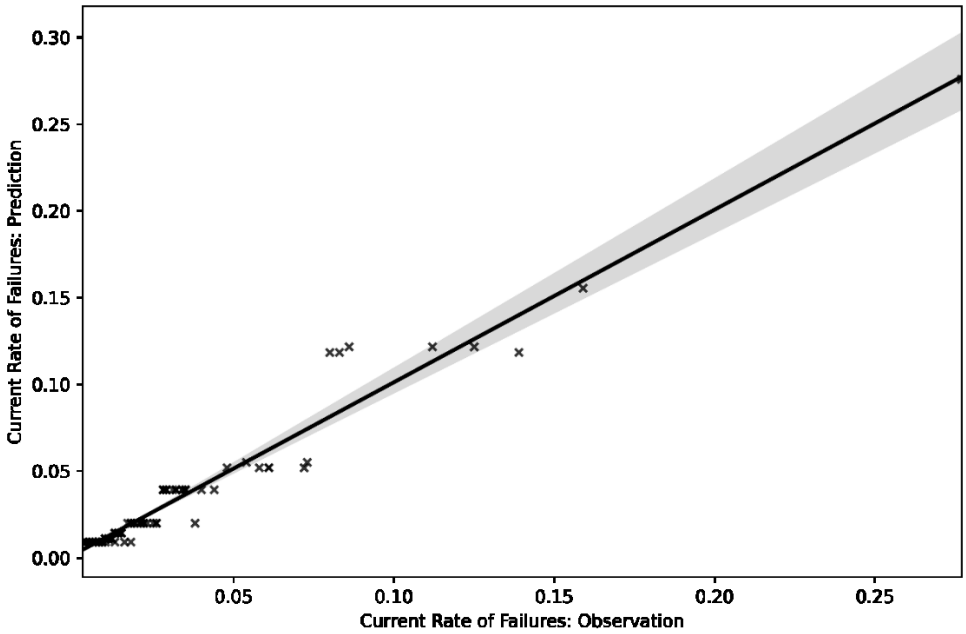


FIGURE 0.60 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – XGBOOST) – ST.JOHN'S

Barrie

- ***Age at First Failure***

Barrie is the final network analyzed in terms of regression. The total number of pipes in the final file is 261, including different input variables such as anode status, soil type, material, diameter, protection status, casing material, restrained, and, more importantly, age at first failure. Cast iron makes up 63.22% of total data points. Meanwhile, ductile iron and PVC are the most frequent type after cast iron, with 24.52% and 7.28% contribution to all pipes with at least one failure. The percentage of other materials can be noticed in the given graph (Figure 0.61).

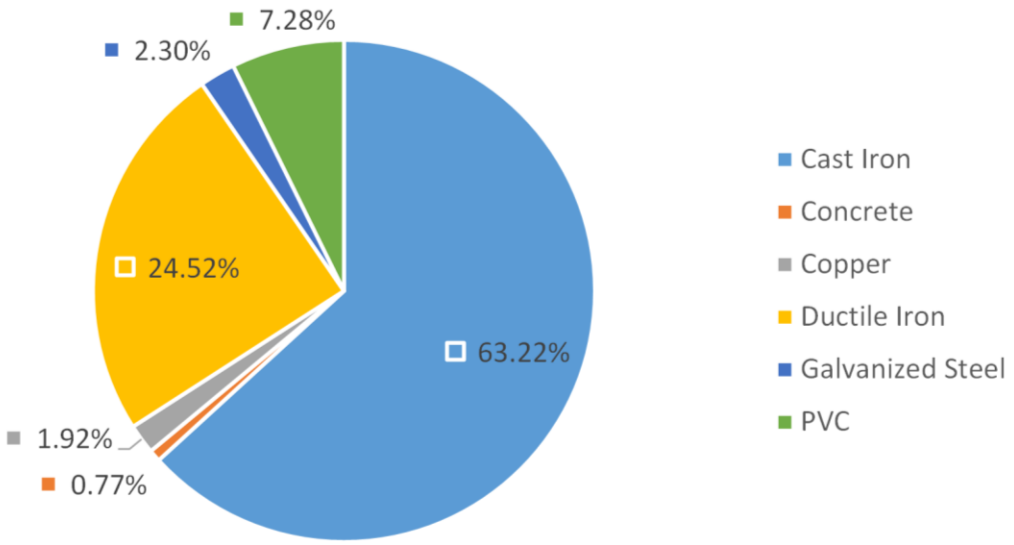


FIGURE 0.61 - PERCENTAGE OF EACH MATERIAL WITHIN REGRESSION ANALYSIS (AGE TO FIRST FAILURE) – BARRIE

The galvanized steel, which accounts for almost 2.30% of total pipes with at least one failure, has the highest average age at the first failure, 51.83 years. The distribution of other materials based on the average age can be seen in the given box plot (Figure 0.62).

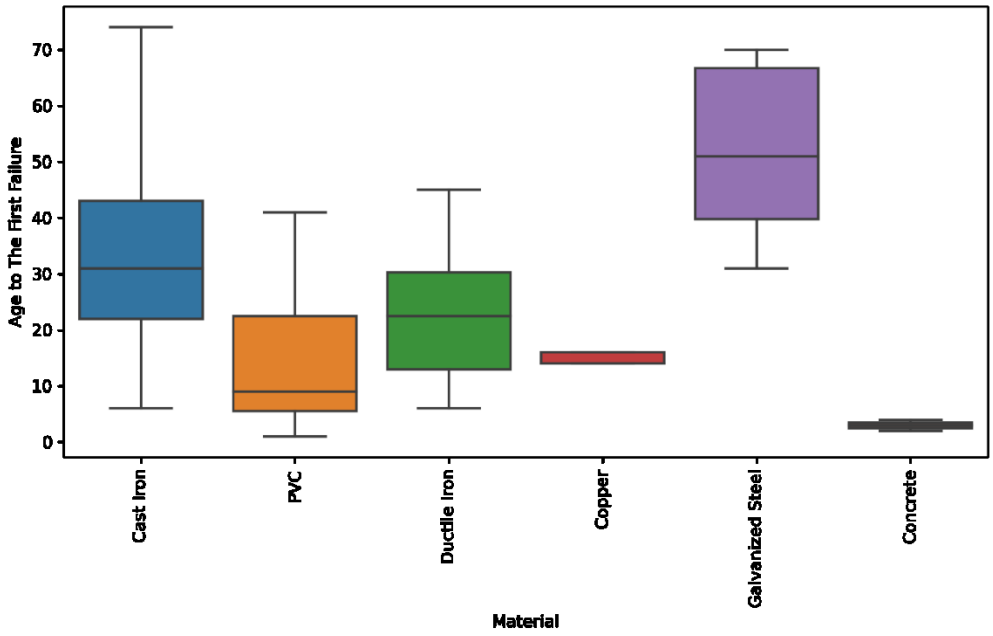


FIGURE 0.62 - DISTRIBUTION OF AGE TO FIRST FAILURE BASED ON TYPE OF MATERIAL (BARRIE)

The provided table shows the final results of the regression analysis for Barrie. Again, random forest demonstrated the best performance for AM category with an RMSE of 14.2 and an R-Squared of 0.35. ANN also showed good accuracy with the R-Squared of 0.32 compared to random forest. For the cast iron group, however, XGBOOST is the best model. The RMSE and R-Squared for XGBOOST are 16.5 and 0.25, in successive (TABLE 0.25).

TABLE 0.250.25 - REGRESSION METRICS (BARRIE)

Algorithm	RMSE		MSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron	AM	Cast Iron
ElasticNet	*14.8*	*18.6*	*218.8*	*345.6*	*0.29*	*0.05*
Random Forest	*14.2*	*17.0*	*202.2*	*289.9*	*0.35*	*0.20*
XGBOOST	*16.0*	*16.5*	*256.3*	*272.9*	*0.17*	*0.25*
ANN	*14.4*	*19.9*	*209.0*	*396.3*	*0.32*	*N.A*

* AM = All Materials

The given regression plot compares the predicted value and actual value of age at the first failure, and it shows that even random forest with the best performance can not find a rational pattern for the Barrie network (Figure 0.63).

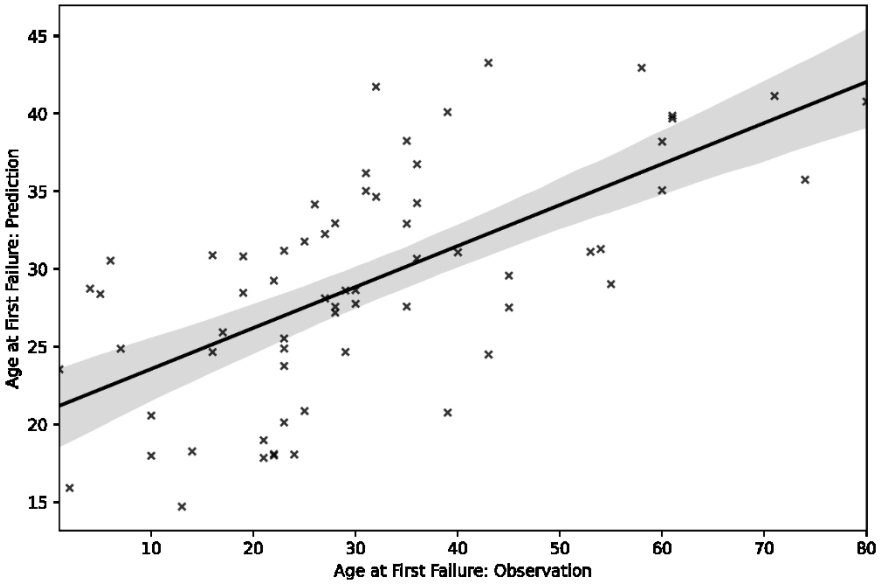


FIGURE 0.63 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – RANDOM FOREST) – BARRIE

- ***Current Rate of Failure***

Regarding the current rate of failure analysis in Barrie, several input variables were used. These variables consist of pipe depth, anode status, service type, material, diameter, protection status, length, casing material, restrained, age, the current rate of failure, and previous rate of failure. Based on the available information, the average current rate of failure is higher during the early-stage bathtub curve. Several peaks can be seen within the given graph, around age 8 and 18. The following graph shows the average current failure rate based on the most recent failure age (Figure 0.64).

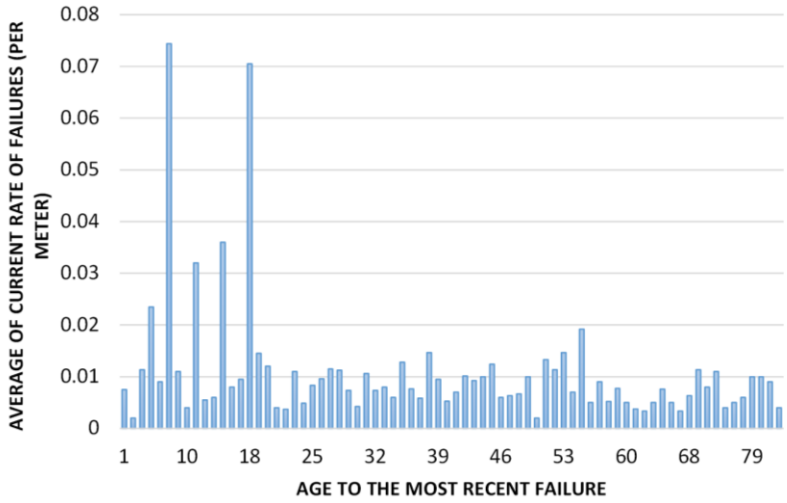


FIGURE 0.64 - AVERAGE OF CURRENT RATE OF FAILURE BASED ON AGE (BARRIE)

The given table shows the result of regression analysis based on the different metrics (TABLE 0.26). As shown, for AM category, XGBOOST indicated a better performance with an R-Squared of 0.60. The accuracy of this model then declined for the cast iron group and dropped to 0.343. The random forest, however, experienced an inverse performance. The accuracy of this model increased from 0.50 for AM category to 0.67 for the cast iron group. This indicates that the performance of each model may vary among different categories, and one uniform model can be employed for all groups of pipes.

TABLE 0.260.26 – REGRESSION METRICS – CURRENT RATE OF FAILURE (BARRIE)

Algorithm	RMSE		R - Squared	
	AM	Cast Iron	AM	Cast Iron
ElasticNet	*0.012*	*0.008*	*N.A*	*0.18*
Random Forest	*0.004*	*0.005*	*0.50*	*0.67*
XGBOOST	*0.003*	*0.007*	*0.60*	*0.343*
ANN	*0.007*	*0.008*	*N.A*	*0.227*

* AM = All Materials

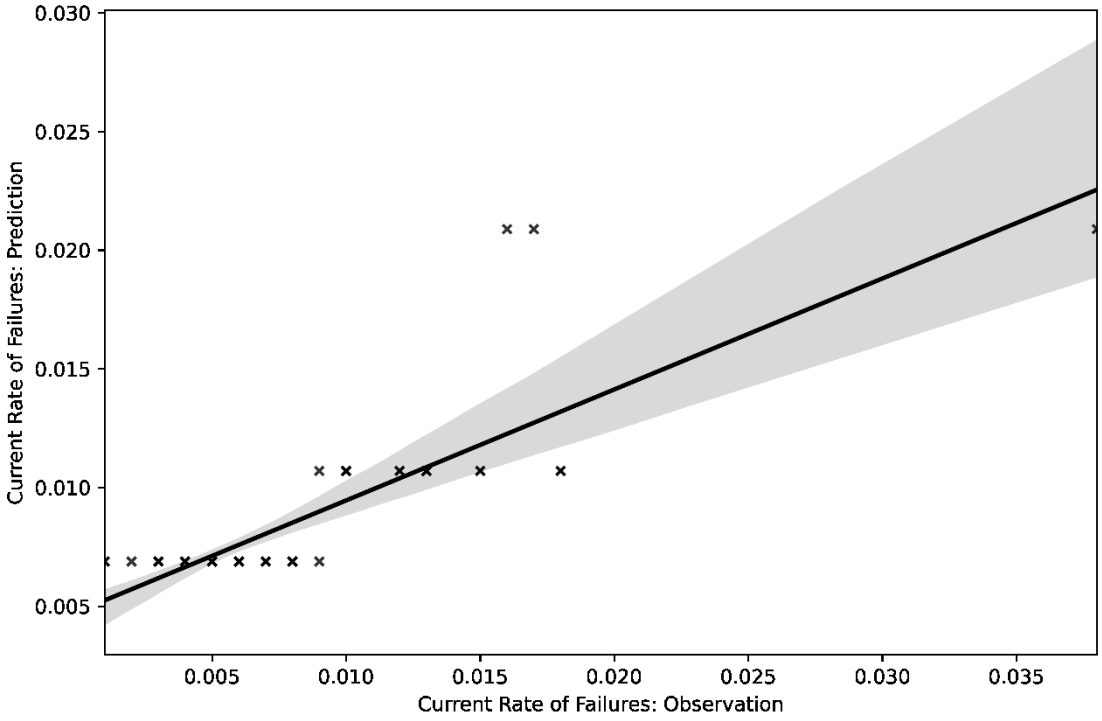


FIGURE 0.65 - REGRESSION PLOT (COMPARE OBSERVATION AND PREDICTION – XGBOOST) – BARRIE

Feature Importance for Regression Models (Age at First Failure)

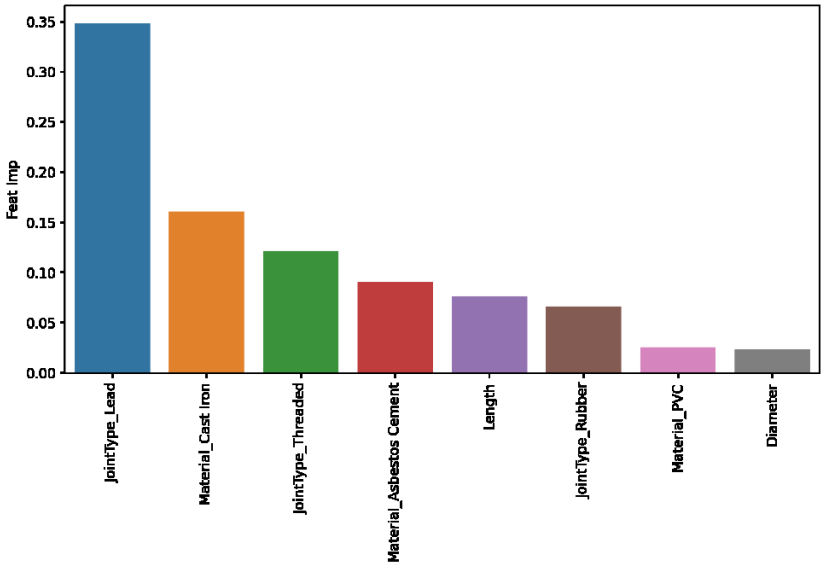


FIGURE 0.66 – FEATURE IMPORTANCE – REGRESSION (AGE AT FIRST FAILURE – SASKATOON)

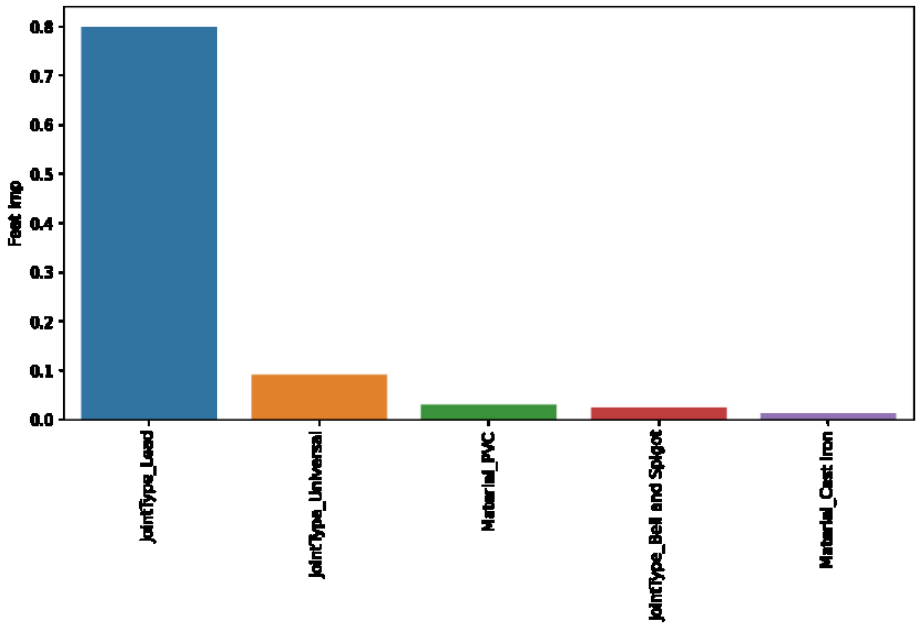


FIGURE 0.67 - FEATURE IMPORTANCE – REGRESSION (AGE AT FIRST FAILURE – WINNIPEG)

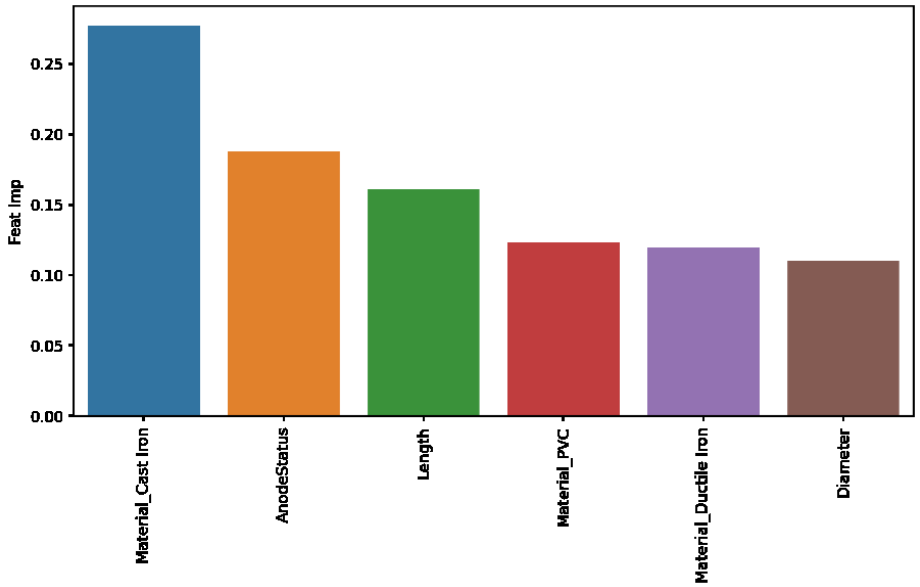


FIGURE 0.68 - FEATURE IMPORTANCE – REGRESSION (AGE AT FIRST FAILURE – KITCHENER)

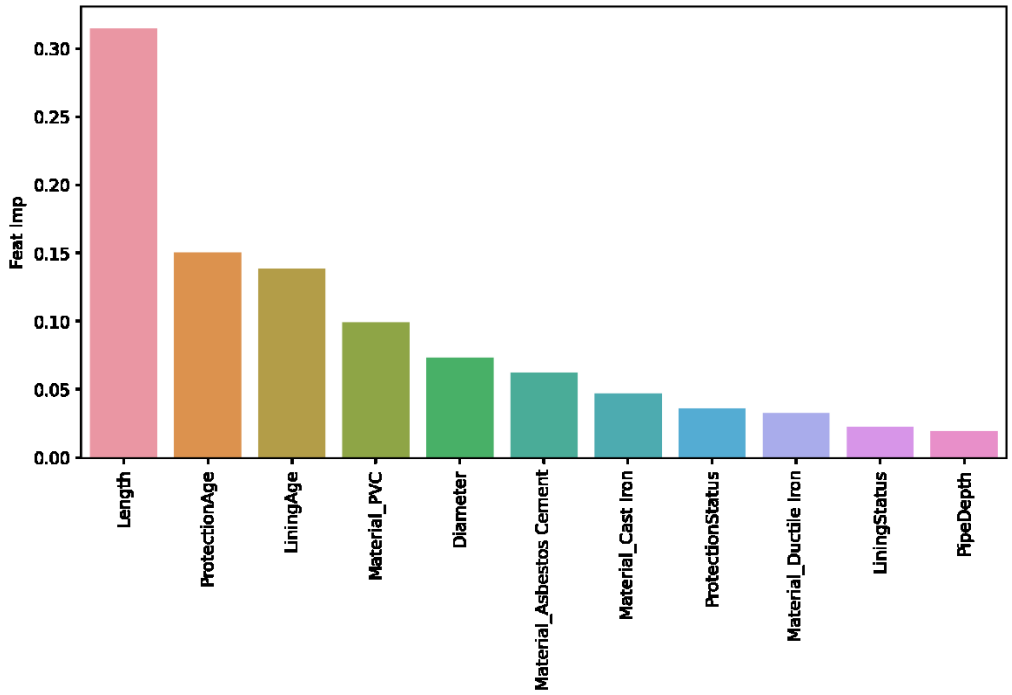


FIGURE 0.69 - FEATURE IMPORTANCE – REGRESSION (AGE AT FIRST FAILURE – MARKHAM)

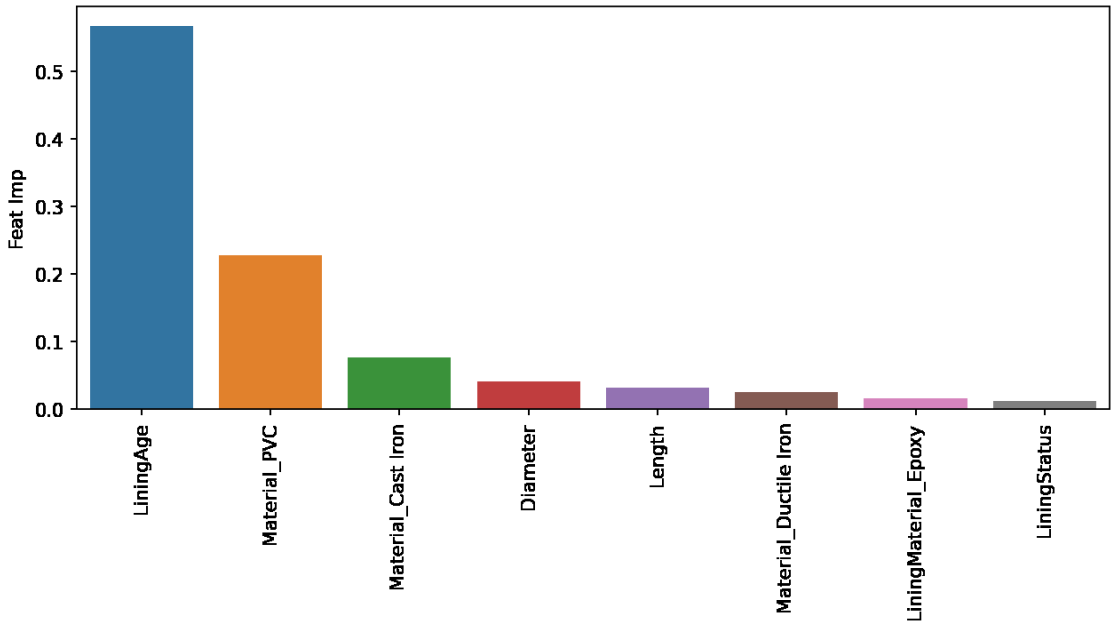


FIGURE 0.70 - FEATURE IMPORTANCE – REGRESSION (AGE AT FIRST FAILURE – WATERLOO)

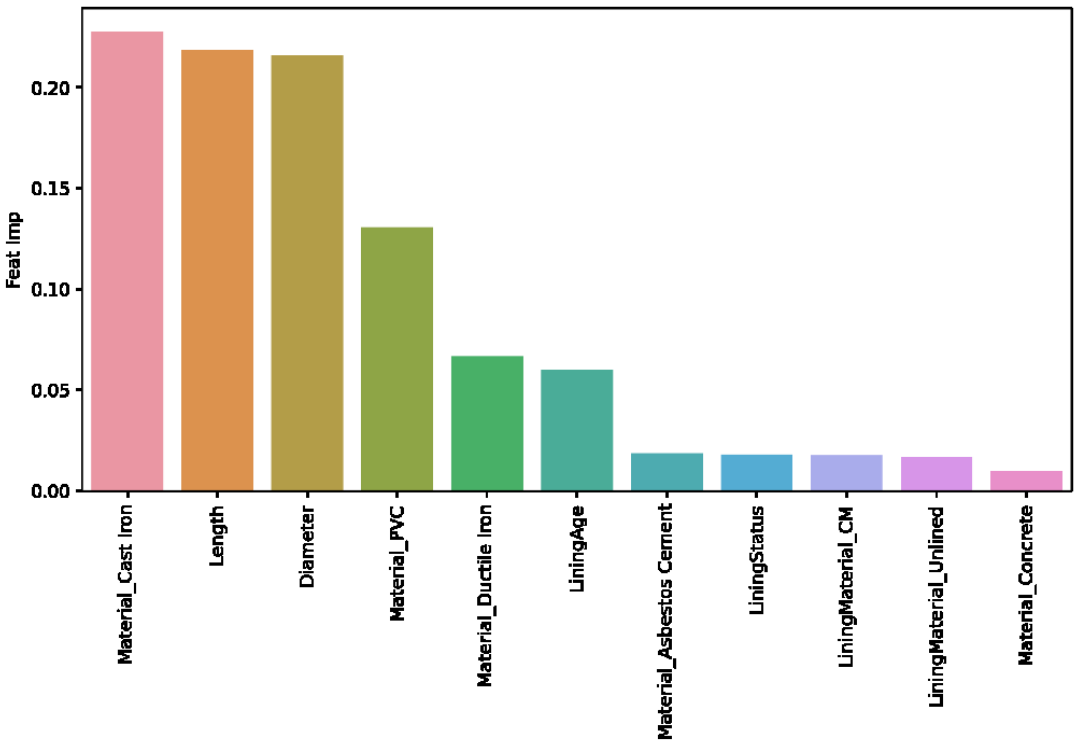


FIGURE 0.71 - FEATURE IMPORTANCE – REGRESSION (AGE AT FIRST FAILURE – REGION OF WATERLOO)

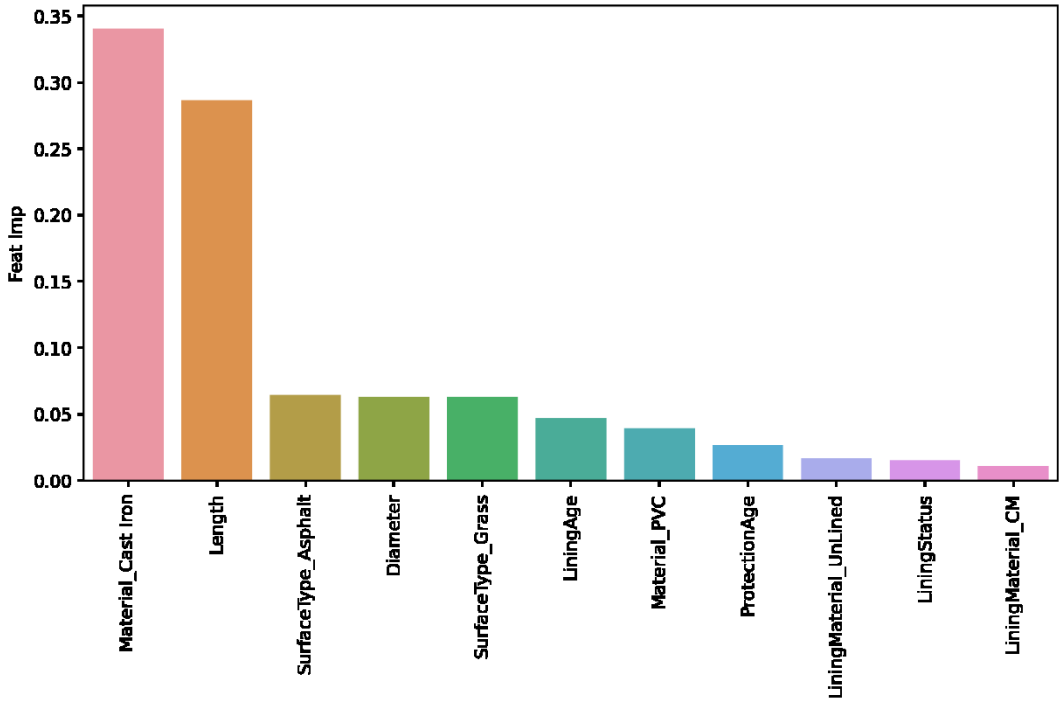


FIGURE 0.72 - FEATURE IMPORTANCE – REGRESSION (AGE AT FIRST FAILURE – REGION OF DURHAM)

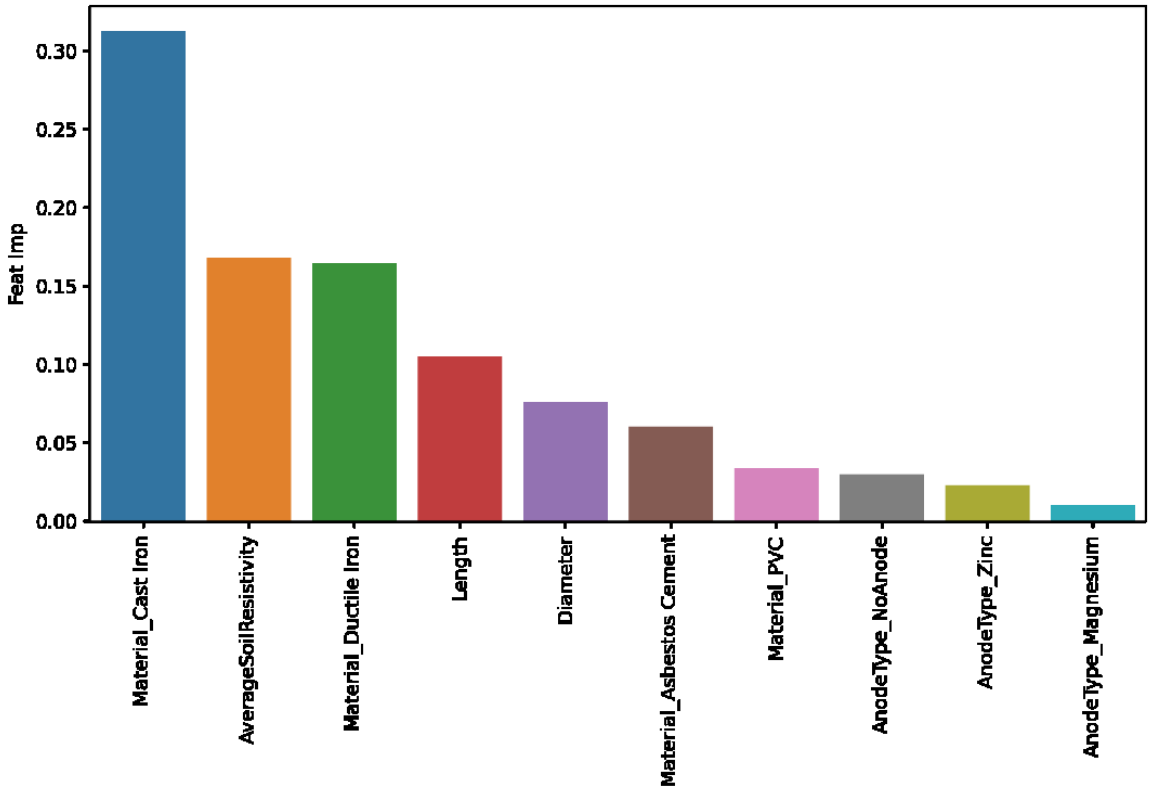


FIGURE 0.73 - FEATURE IMPORTANCE – REGRESSION (AGE AT FIRST FAILURE – CALGARY)

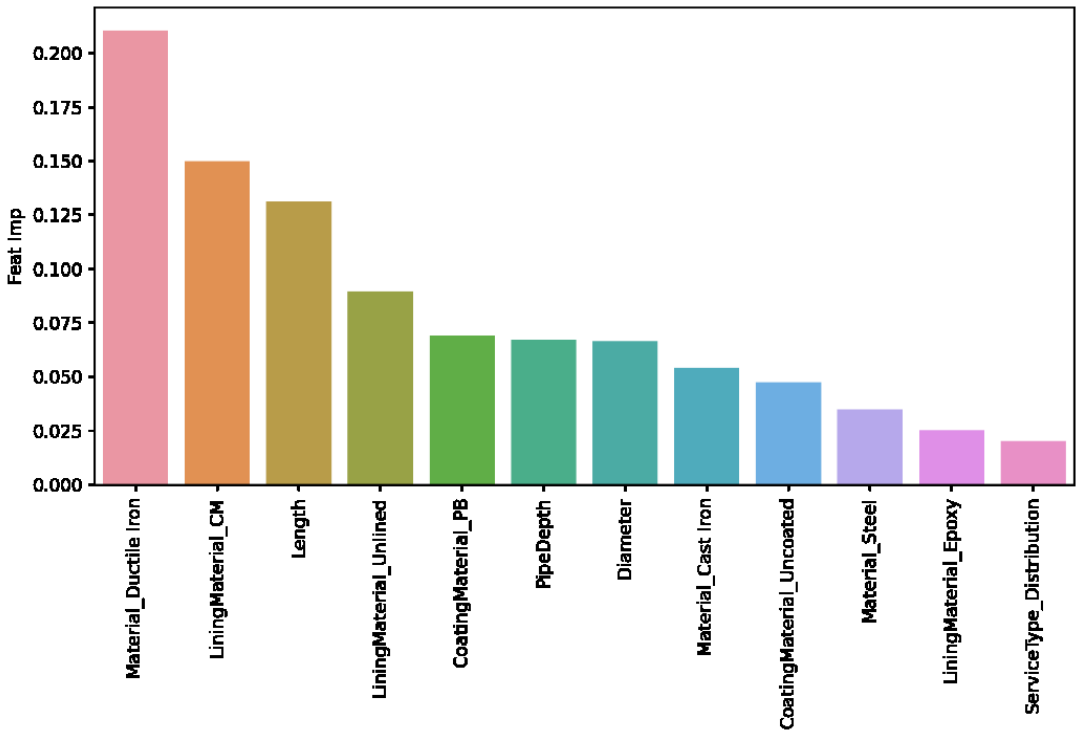


FIGURE 0.74 - FEATURE IMPORTANCE – REGRESSION (AGE AT FIRST FAILURE – VANCOUVER)

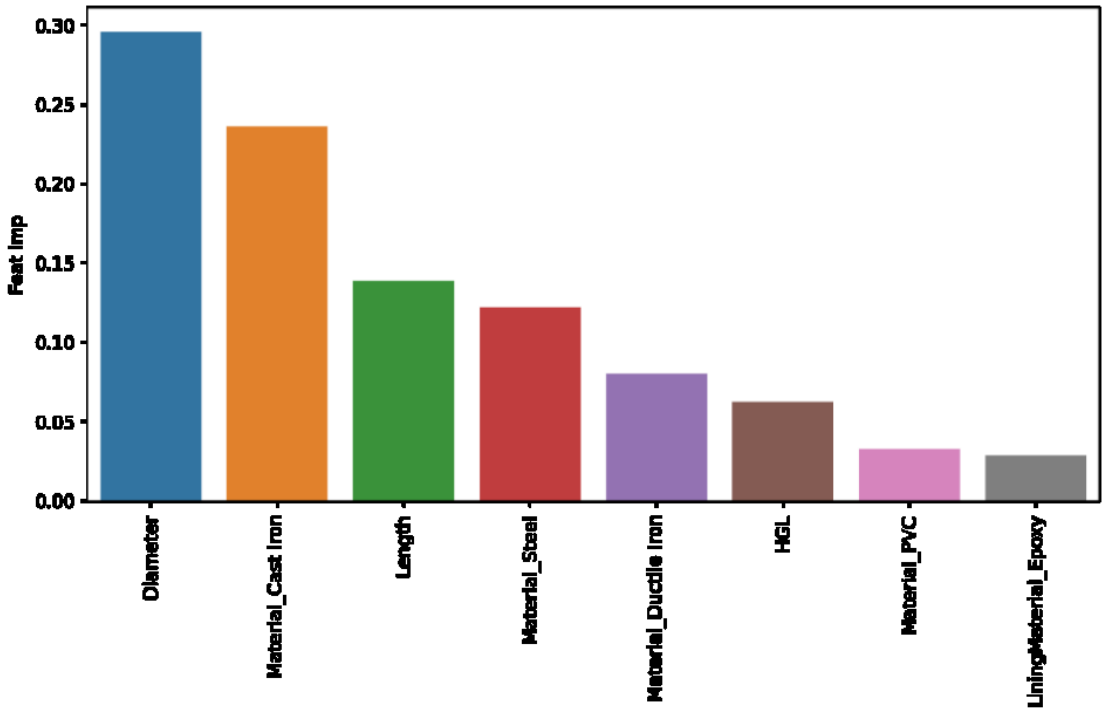


FIGURE 0.75 - FEATURE IMPORTANCE – REGRESSION (AGE AT FIRST FAILURE – VICTORIA)

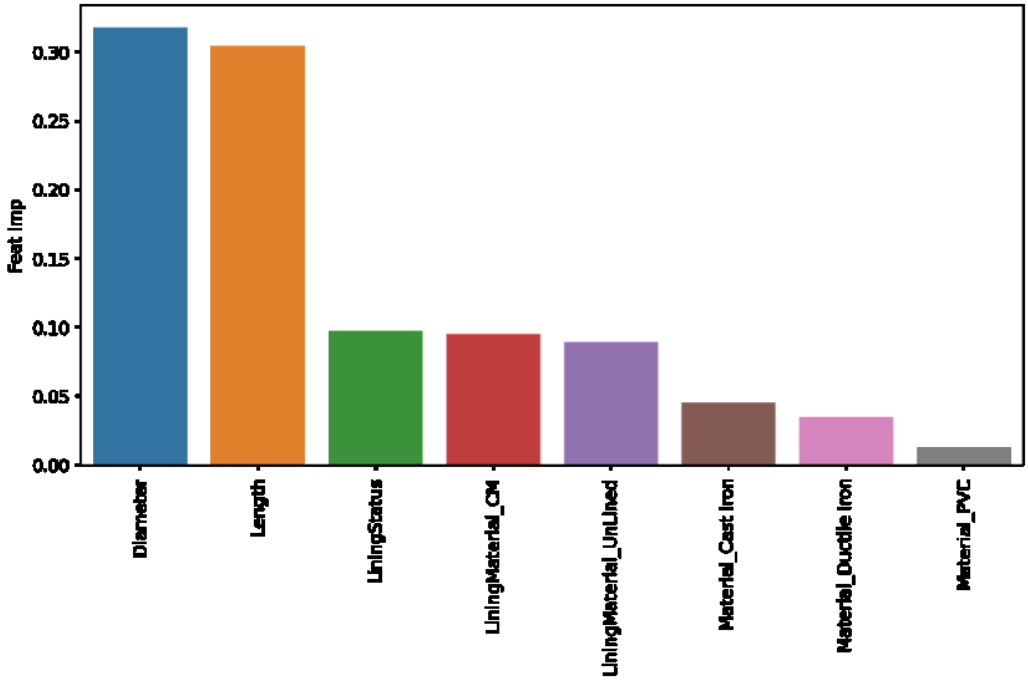


FIGURE 0.76 - FEATURE IMPORTANCE – REGRESSION (AGE AT FIRST FAILURE – HALIFAX)

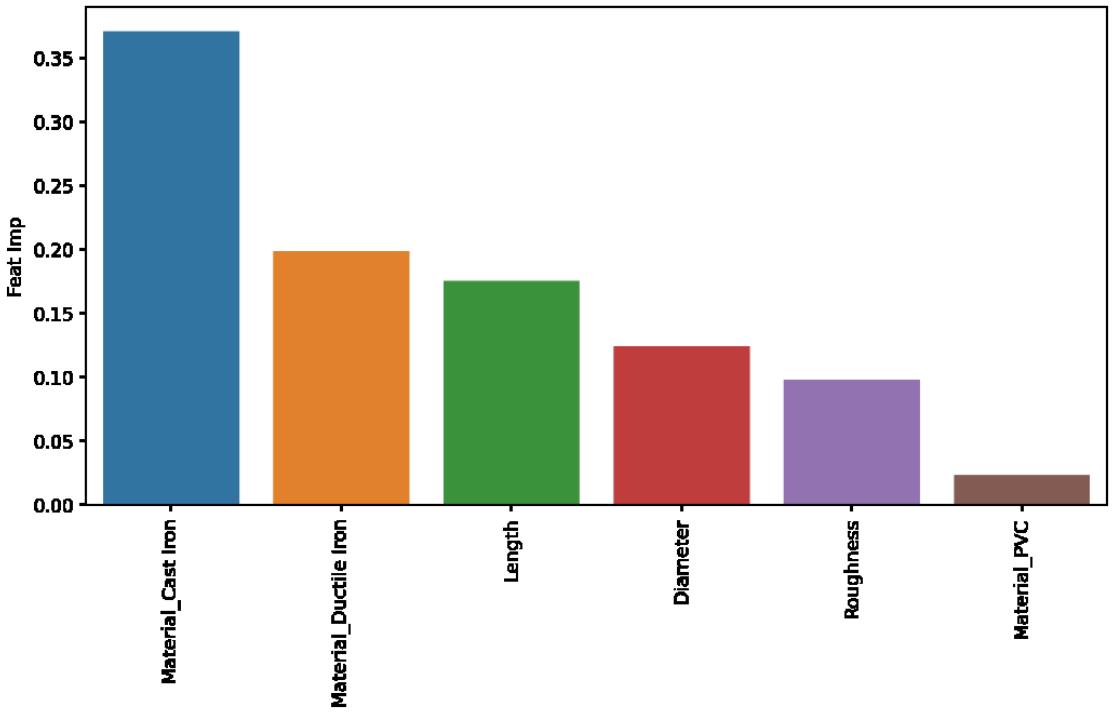


FIGURE 0.77 - FEATURE IMPORTANCE – REGRESSION (AGE AT FIRST FAILURE – ST. JOHN'S)

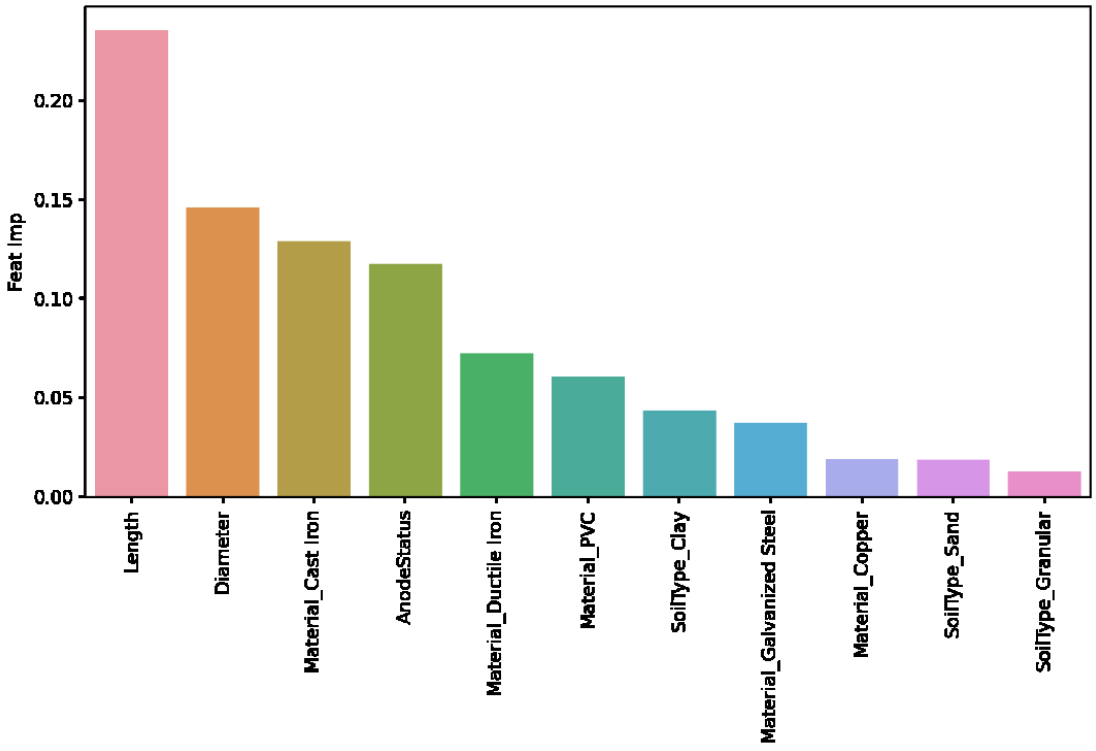


FIGURE 0.78 - FEATURE IMPORTANCE – REGRESSION (AGE AT FIRST FAILURE – BARRIE)

APPENDIX F – CORRELATION MATRIX (SPEARMAN – CLASSIFICATION MODELS)

Spearman correlation analysis has been conducted to find any significant correlation between numerical attributes. For example, for Saskatoon, the given correlation matrix indicates the correlation score between age at the first failure and other attributes for broken pipes (class 1). For asbestos cement pipes and threaded joint types, there are correlation scores of -0.42 and -0.43, respectively. This negative correlation shows that when joint type is threaded or pipe is asbestos cement, the age at the first failure is lower than in other situations. However, this correlation is not considered as strong since its value is below -0.5.

On the other hand, the correlation scores for cast iron pipes and lead joint are positive, with the score 0.49 and 0.52, respectively, indicating that the age at the first failure is higher whenever these two features are available. Correlation scores for other materials and other attributes can be found in the given matrix (Figure 0.1). It should be noted that this matrix only belongs to the broken pipes.

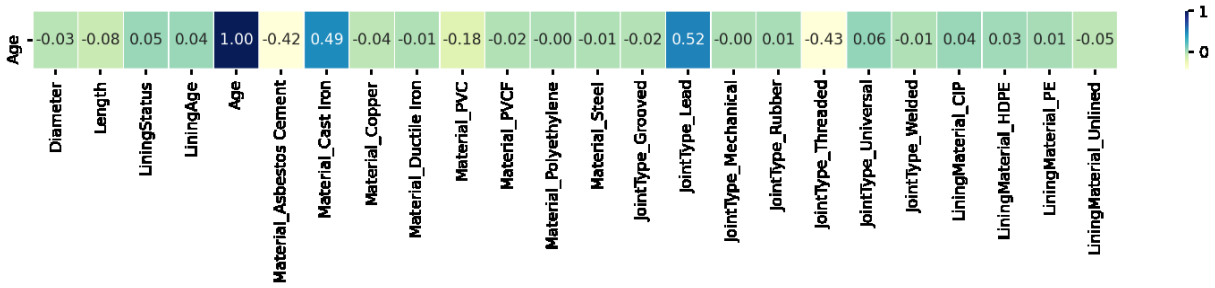


Figure 0.1 – Spearman correlation analysis for class 1 or broken pipes (Saskatoon)

The Spearman analysis was also applied to Winnipeg`s classification dataset, and no significant correlation was found between the numerical attributes (Figure 0.2). Among broken pipes, only cast iron is somewhat correlated with age at the first failure, with a score of 0.52, showing that age at first break increases where cast iron does exist. The given figure shows how attributes within the Calgary network for broken pipes are correlated.

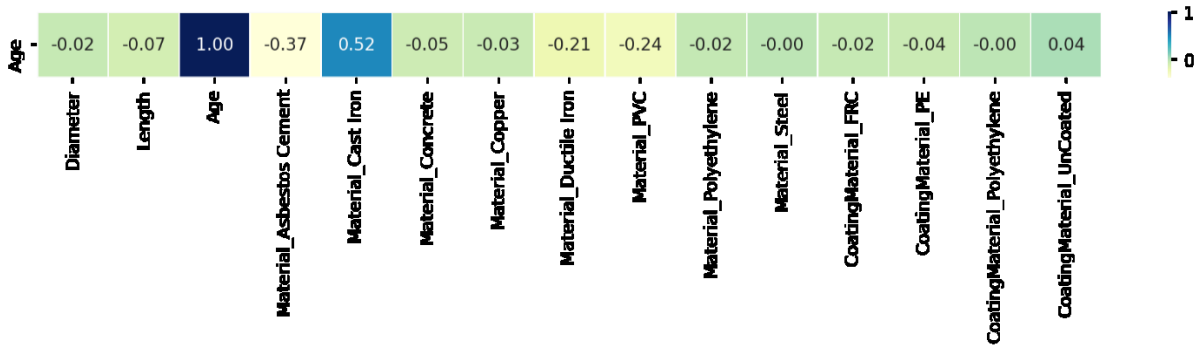


Figure 0.2 - Spearman correlation analysis for class 1 or broken pipes (Winnipeg)

For Kitchener, from the correlation matrix that has been produced based on Spearman Correlation Analysis, there is no significant correlation between input variables and the age at the first failure (Figure 0.3). The highest positive correlation is related to the cast iron material with the value of 0.52, which somewhat correlates with age at the first failure. This value shows that when cast iron is installed, the age at the first failure increases, although this does not show any specific causations. Moreover, ductile iron with a score of -0.45 seems to have a weak inverse correlation with age at the first failure.

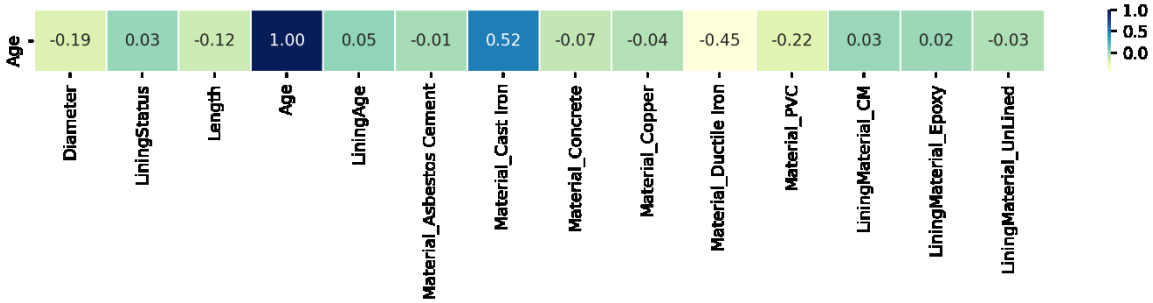


Figure 0.3 – Spearman correlation analysis for class 1 or broken pipes (Kitchener)

In Markham, Spearman Correlation Analysis indicated no significant correlation between age at the first failure and other input variables in the classification dataset. Nonetheless, a weak correlation between PVC and age at the first failure can be noticed, with a value of -0.27. Therefore, the given matrix is recommended to better understand the correlation between age at the first failure and other attributes (Figure 0.4). Same as before, this matrix is created based on only broken pipes.

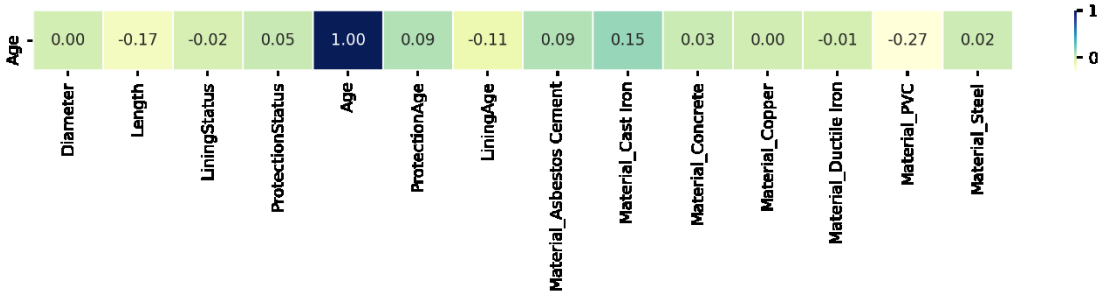


Figure 0.4 - Spearman correlation analysis for class 1 or broken pipes (Markham)

The followings are the correlation matrixes prepared for other utilities in this study.

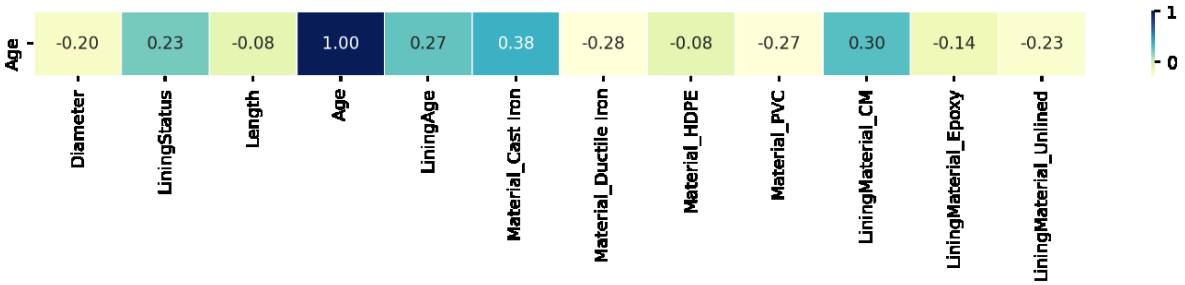


Figure 0.5 - Spearman correlation analysis for class 1 or broken pipes (waterloo)

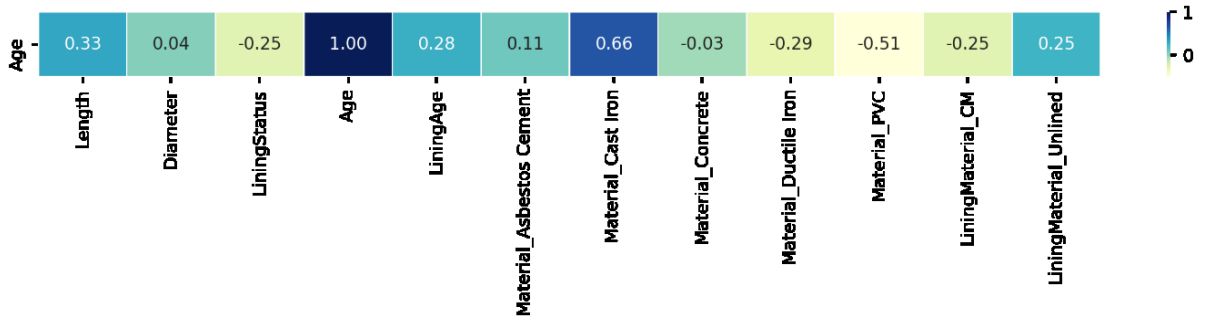


Figure 0.6 - Spearman correlation analysis for class 1 or broken pipes (Region of waterloo)

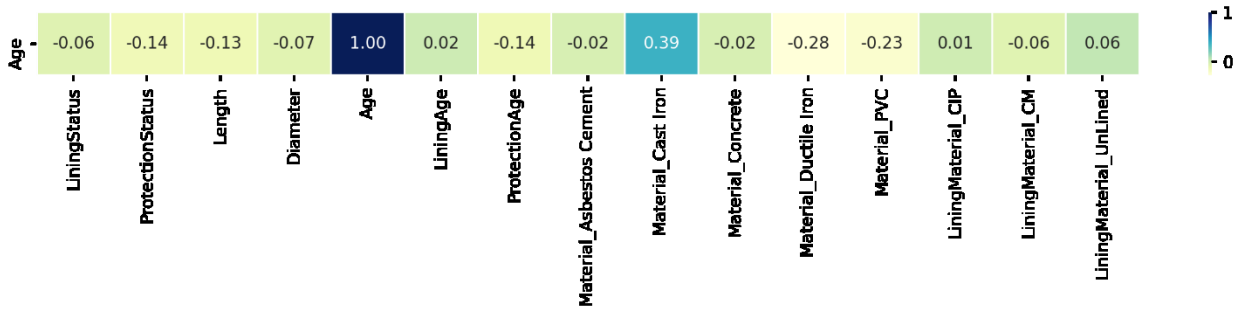


FIGURE 0.7 - SPEARMAN CORRELATION ANALYSIS FOR CLASS 1 OR BROKEN PIPES (REGION OF DURHAM)

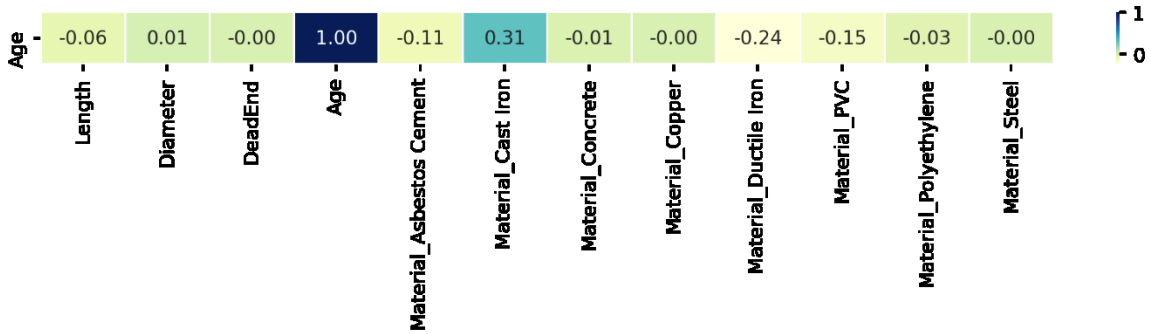


FIGURE 0.8 - SPEARMAN CORRELATION ANALYSIS FOR CLASS 1 OR BROKEN PIPES (CALGARY)

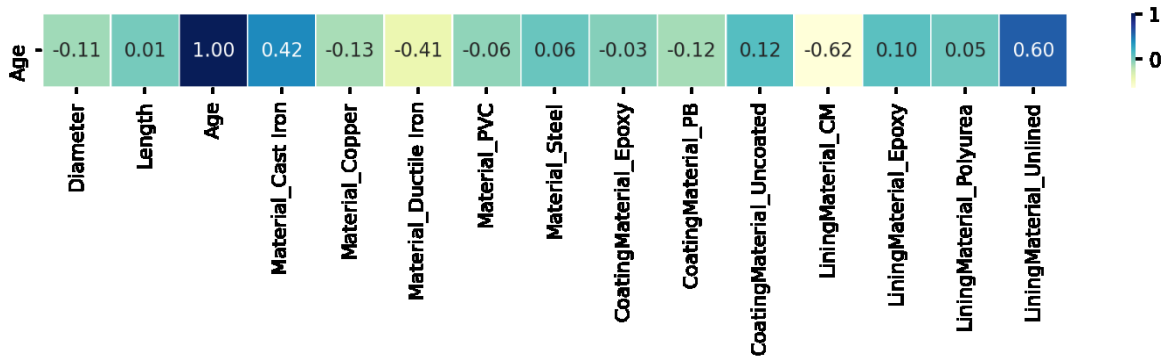


FIGURE 0.9 – SPEARMAN CORRELATION ANALYSIS FOR CLASS 1 OR BROKEN PIPES (VANCOUVER)

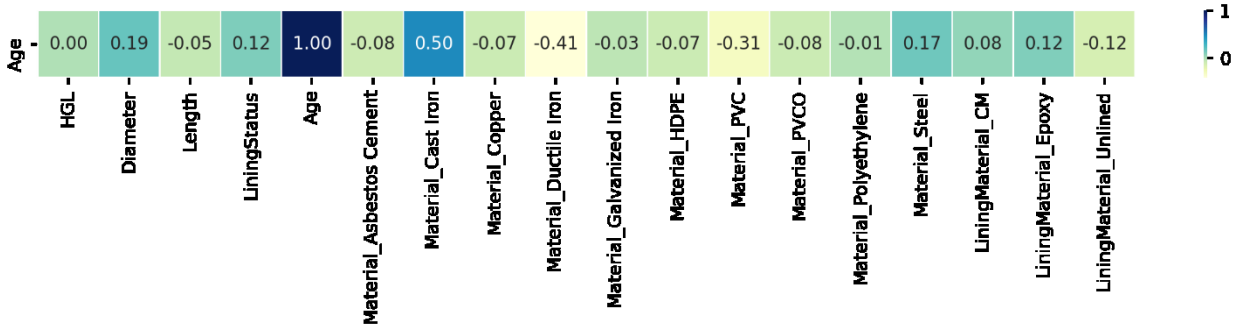


FIGURE 0.10 - SPEARMAN CORRELATION ANALYSIS FOR CLASS 1 OR BROKEN PIPES (VICTORIA)

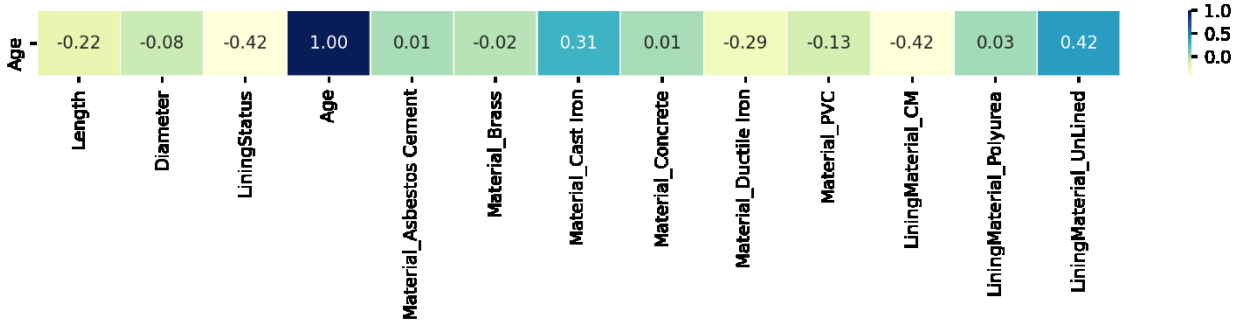


FIGURE 0.11 - SPEARMAN CORRELATION ANALYSIS FOR CLASS 1 OR BROKEN PIPES (HALIFAX)

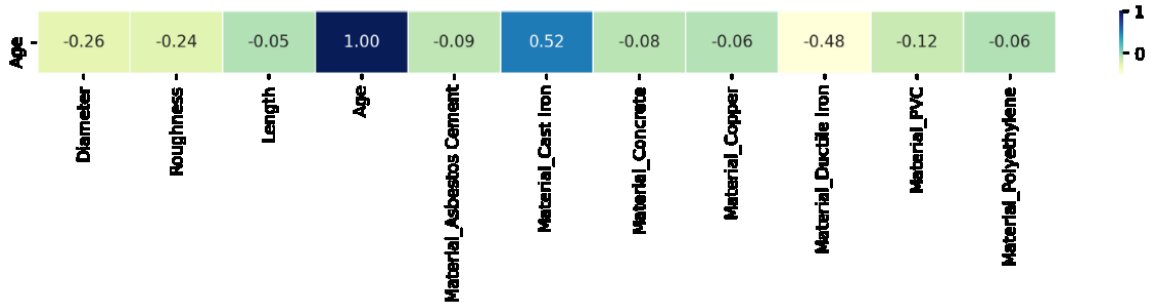


FIGURE 0.12 - SPEARMAN CORRELATION ANALYSIS FOR CLASS 1 OR BROKEN PIPES (ST. JOHN'S)

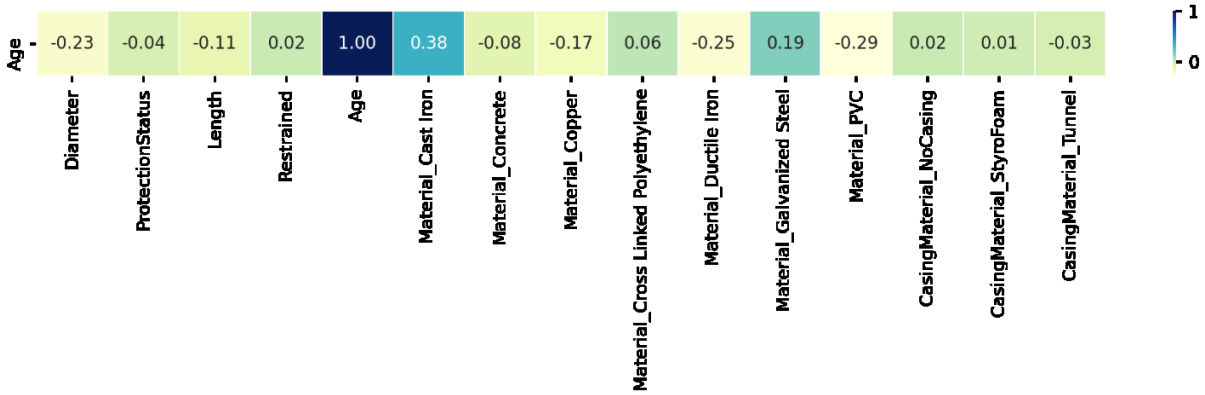


FIGURE 0.13 - SPEARMAN CORRELATION ANALYSIS FOR CLASS 1 OR BROKEN PIPES (BARRIE)

APPENDIX G – CORRELATION MATRIX (SPEARMAN – REGRESSION MODELS – AGE AT FIRST FAILURE)

Spearman correlation analysis results are similar to the classification problem since they consider age at first failure. However, due to differences in the structure of both datasets, the scores for correlation vary a little bit. Given matrix depicts that whenever a joint type is a lead, the number of years to the first failure increases, with a score of 0.56 (Figure 0.1). Same pattern for cast iron material with a smaller score, which is 0.48. These scores can be related to a weak correlation.

On the other hand, when the material is asbestos cement or the joint type is threaded, age at the first failure tends to decline. This can show that these factors may have an adverse effect on the pipe deterioration process.

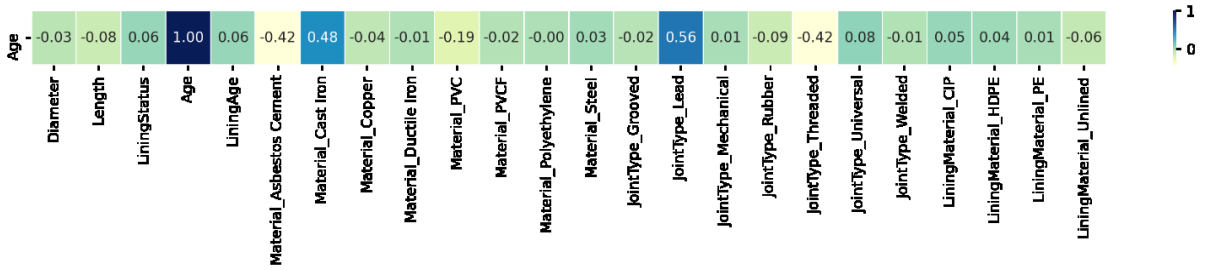


Figure 0.1 – Spearman correlation – Age to first failure (Regression – Saskatoon)

Analyzing the correlation between different attributes with the Spearman analysis shows a strong correlation between age and first failure and lead joins, with an uphill score of 0.64. Cast iron pipe also indicates a positive correlation, although moderate, and the score is 0.54.

Like Saskatoon, asbestos cement negatively correlates with age, indicating that this pipe has experienced its first failure at a younger age. Moreover, the Collar joint also shows a similar pattern. The score for asbestos cement pipe and collar joint are -0.41 and -0.40, respectively. More correlation scores can be found in the given matrix below (Figure 0.2).

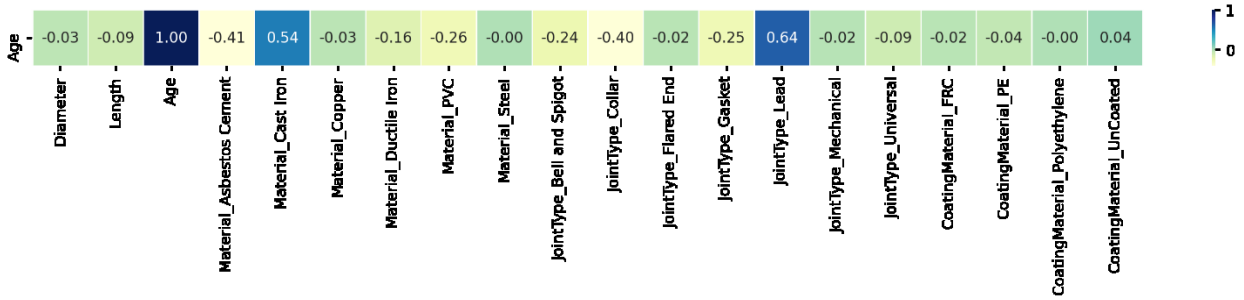


Figure 0.2 - Spearman correlation – Age to first failure (Regression – Winnipeg)

The given figure is prepared based on the Spearman correlation analysis. Cast iron and ductile iron pipes have a positive and negative correlation with age, with a score of 0.52 and -0.45, respectively. In this case, anode status has a weak positive correlation with age at first failure, with a score of 0.37. PVC pipes show a very weak correlation with age in this network (Figure 0.3).

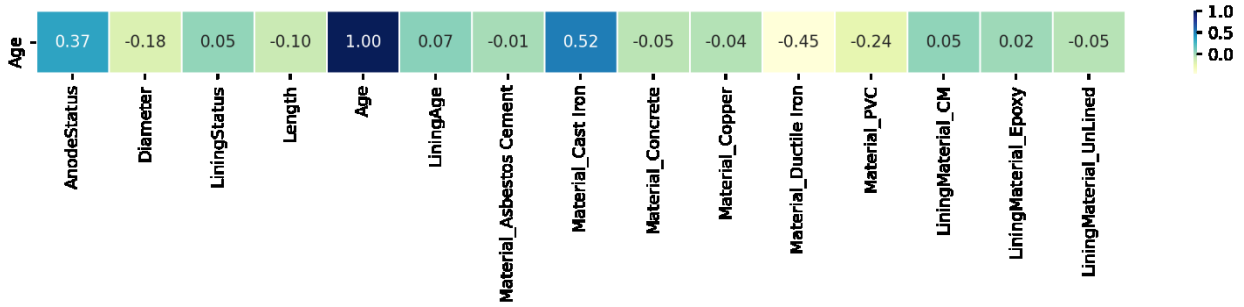


Figure 0.3 - Spearman correlation – Age to first failure (Regression – Kitchener)

No significant correlation was found between age at first failure and the available attributes for the Markham network. The only noticeable correlations belong to PVC pipe and length, with a score of -0.25 and -0.16, respectively. However, these scores are considered a very weak correlation and do not show anything specific (Figure 0.4).

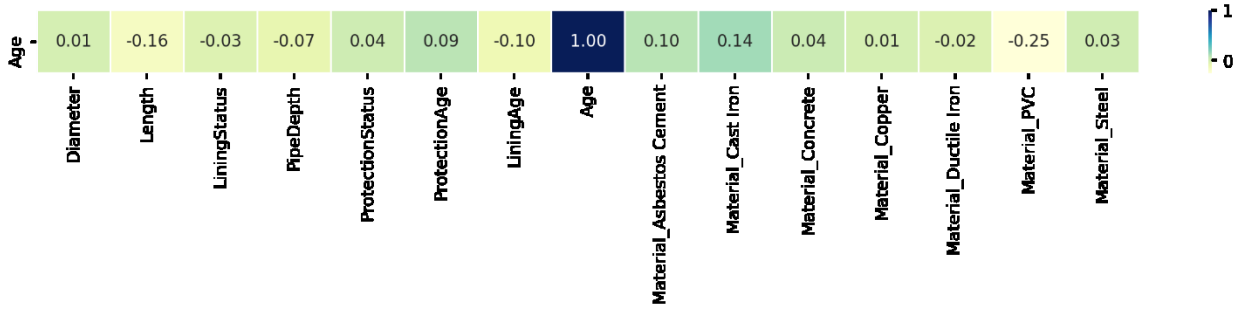


Figure 0.4 - Spearman correlation – Age to first failure (Regression – Markham)

Figure 0.5 shows the results of the Spearman correlation analysis. With a score of 0.38, the cast iron pipe has a weak uphill correlation with age at first failure, which is insignificant. Ductile iron and PVC also negatively correlated with age, -0.28 and -0.27, in successive.

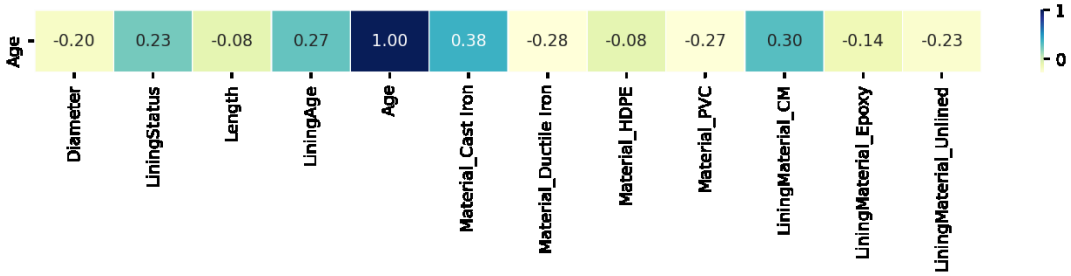


Figure 0.5 - Spearman correlation – Age to first failure (Regression – Waterloo)

The Spearman correlation analysis found a strong correlation between age at first failure and cast iron pipes, with a score of 0.65. This score indicates the increase in the age at first failure when cast iron pipes exist. Conversely, PVC and ductile iron pipes show a weak negative

correlation with age at first failure. The score for PVC pipe is -0.48 and for ductile iron -0.34 (Figure 0.6).

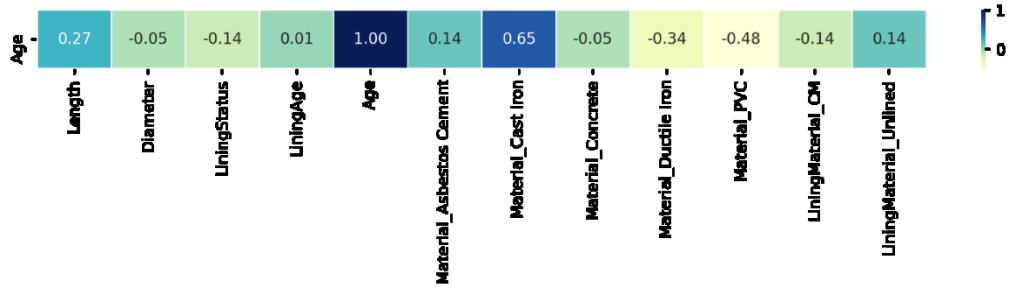


Figure 0.6 - Spearman correlation – Age to first failure (Regression – Region of Waterloo)

Durham`s attributes did now show any significant correlation with age at first failure. Nonetheless, there is a weak upward correlation between age and cast-iron pipe, which is 0.32. Correlation scores for other variables are given in the given figure (Figure 0.7).

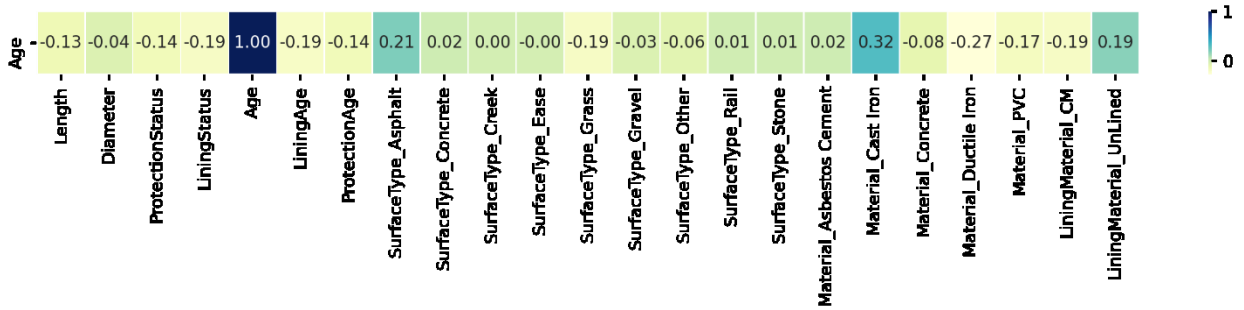


FIGURE 0.7 - SPEARMAN CORRELATION – AGE TO FIRST FAILURE (REGRESSION – REGION OF DURHAM)

Correlation analysis was applied to the Calgary regression dataset. The given matrix shows no marked correlation between age at first failure and input variables (Figure 0.8). Nevertheless, cast iron with a score of 0.28 has the highest correlation with age.

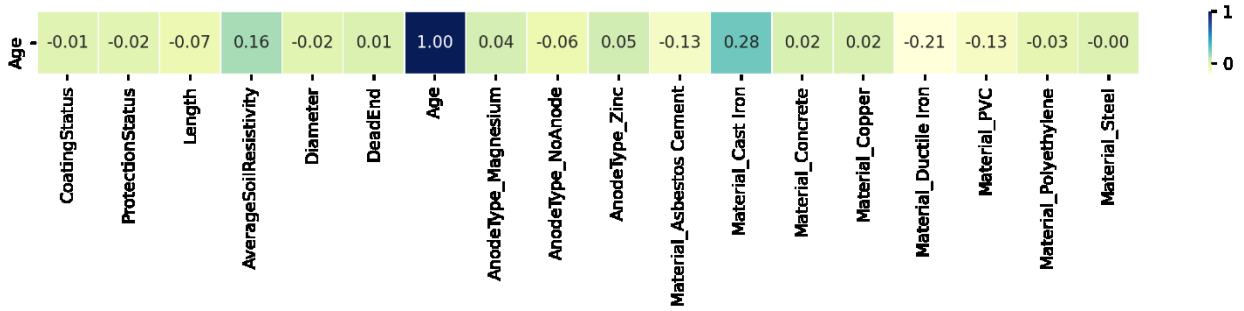


Figure 0.8 - Spearman correlation – Age to first failure (Regression – Calgary)

Cement mortar lining seems to have a moderate downward correlation score with age at first failure in the Vancouver network, with a score of -0.57. Therefore, the effect of this variable should be analyzed carefully for the network, and it may require in-site investigation. Conversely, unlined pipes have a moderate positive correlation with age at first failure, 0.54.

In terms of material, cast iron and ductile iron have a positive and negative correlation, with a score of 0.30 and -0.40, respectively. Correlation scores for other attributes can be found in the given matrix (Figure 0.9).

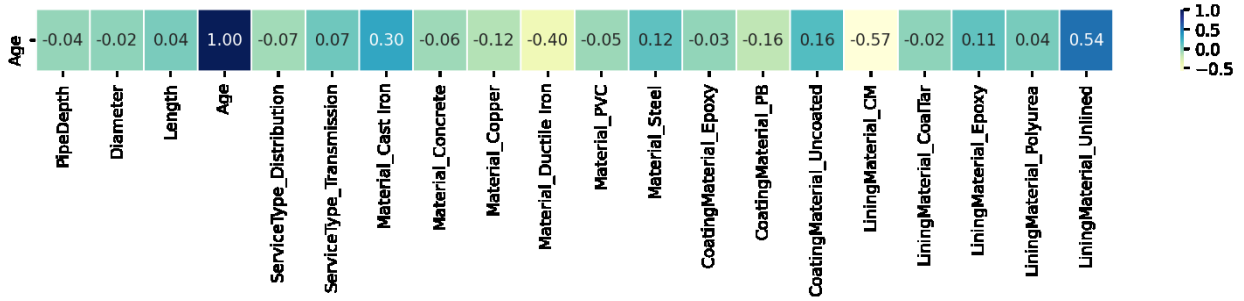


Figure 0.9 - Spearman correlation – Age to first failure (Regression – Vancouver)

A weak correlation was found between age at first failure and other variables in the Victoria network. However, it is worth mentioning that cast iron with a positive correlation of 0.46 has the most considerable correlation with the years to the first failure. Ductile iron also has a moderate correlation of -0.43 with the target of this study. As mentioned before, other attributes have no or very weak correlation with the dependent variable (Figure 0.10).

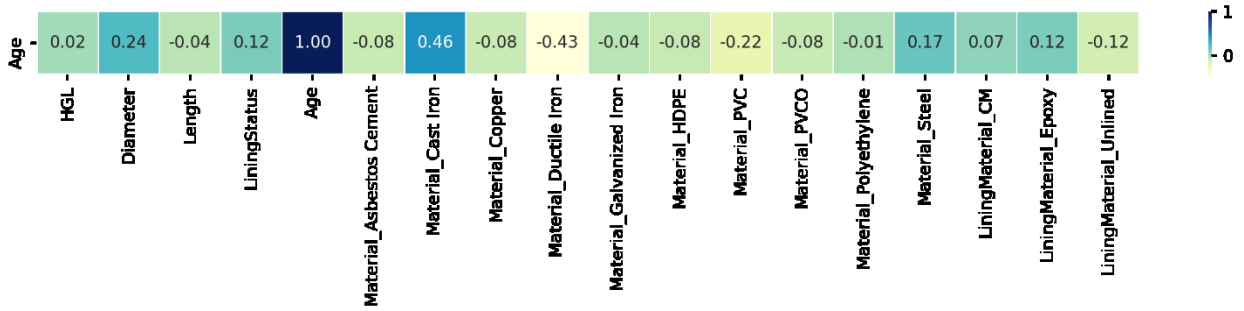


FIGURE 0.10 - SPEARMAN CORRELATION – AGE TO FIRST FAILURE (REGRESSION – VICTORIA)

The given figure reveals that lining status has a moderate negative correlation with age at the first failure, with a score of -0.42 (Figure 0.11). However, comparing the results for Halifax with other utilities indicated that lining conditions should be investigated more carefully and in more detail. Cement mortal, for instance, with a moderate correlation score of -0.42, is among the attributes with the highest correlation. Regarding material, cast iron and ductile iron have weak scores of 0.31 and -0.29, respectively.

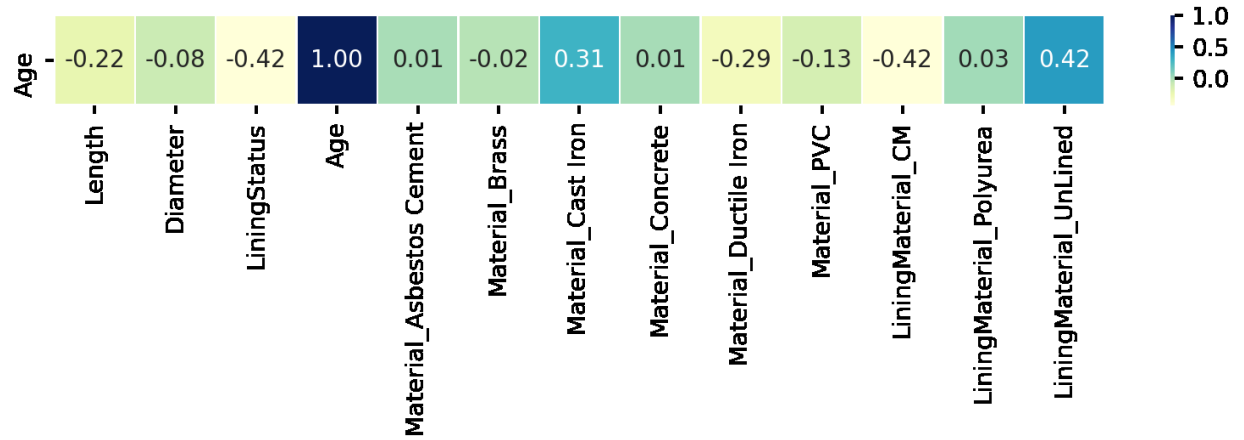


Figure 0.11 - Spearman correlation – Age to first failure (Regression – Halifax)

Based on the Spearman correlation analysis, some materials were found to have a moderate correlation with age at first failure. For instance, cast iron pipe with a score of 0.52 indicated the highest correlation with age, followed by ductile iron with a negative correlation of -0.48. Furthermore, other input variables such as diameter and material indicated a very weak correlation score of -0.26 and -0.24, respectively (Figure 0.12).

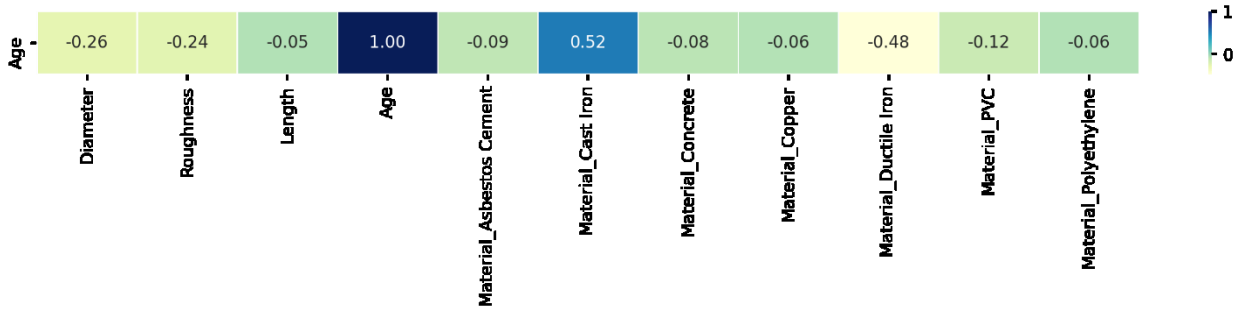


Figure 0.12 - Spearman correlation – Age to first failure (Regression – St. John's)

Most of the attributes with Barrie seem to have a weak correlation with age at the first failure. For instance, cast iron pipe indicated the highest correlation score, 0.40, with age at the first failure, followed by PVC with a weak downward score of -0.30. The correlation scores for other attributes can be seen in the given figure (Figure 0.13).

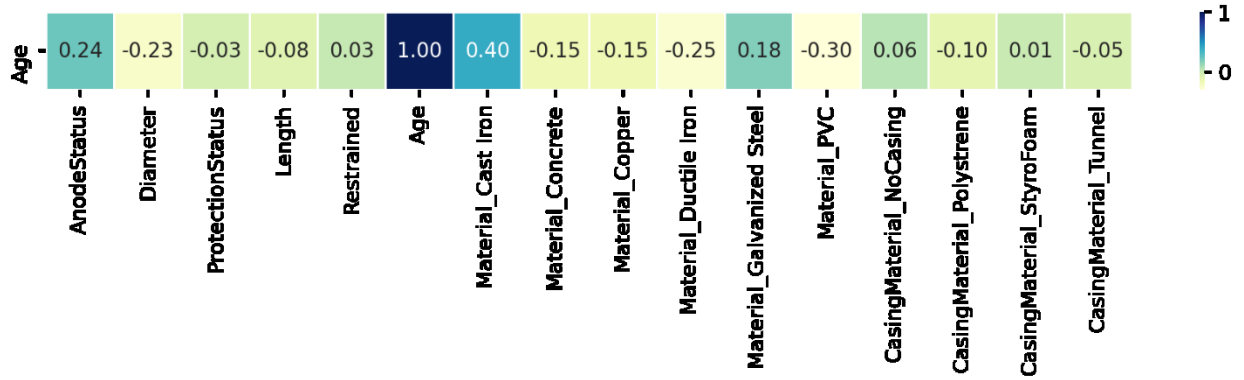


FIGURE 0.13 - SPEARMAN CORRELATION – AGE TO FIRST FAILURE (REGRESSION – BARRIE)

APPENDIX H – COMPARING ALL CLASSIFICATION RESULTS

COMPARISON	Random Forest				XGBOOST				Logisitic Regression				ANN			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
All Materials	90%	94%	35%	52%	90%	88%	40%	55%	88%	73%	32%	45%	88%	90%	21%	35%
Cast Iron	88%	90%	56%	69%	87%	81%	62%	70%	85%	84%	46%	59%	87%	86%	56%	68%
HDPE	**Not Enough Information, Only One Broken Pipe**															
PVC	96%	100%	50%	67%	96%	100%	50%	67%	81%	25%	75%	38%	92%	50%	25%	33%
Ductile Iron	95%	0%	0%	0%	93%	0%	0%	0%	74%	12%	62%	20%	93%	20%	8%	11%
Cluster 0	87%	91%	48%	63%	86%	78%	55%	64%	85%	78%	48%	60%	87%	85%	53%	65%
Cluster 1	95%	100%	6%	11%	95%	100%	11%	20%	70%	11%	61%	18%	92%	9%	6%	7%
Cluster 2	**Not Enough Information, Only One Broken Pipe**															
Cluster 3	75%	0%	0%	0%	67%	0%	0%	0%	67%	0%	0%	0%	67%	0%	0%	0%

FIGURE 0.1 – COMPARING RESULTS FOR DIFFERENT GROUPS (VICTORIA – CLASSIFICATION MODELS)

COMPARISON	Random Forest				XGBOOST				Logisitic Regression				ANN			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
All Materials	98%	100%	4%	8%	97%	20%	4%	7%	71%	6%	70%	11%	97%	38%	13%	19%
Cast Iron	92%	80%	44%	57%	87%	50%	22%	31%	72%	24%	56%	33%	86%	43%	33%	38%

FIGURE 0.2 - COMPARING RESULTS FOR DIFFERENT GROUPS (REGION OF WATERLOO – CLASSIFICATION MODELS)

COMPARISON	Random Forest				XGBOOST				Logisitic Regression				ANN			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
All Materials	97%	84%	62%	72%	97%	77%	66%	71%	97%	83%	57%	68%	97%	82%	66%	73%
Cast Iron	85%	89%	82%	85%	82%	84%	82%	83%	85%	86%	84%	85%	78%	79%	79%	79%
Copper	92%	0%	0%	0%	92%	33%	50%	40%	68%	0%	0%	0%	97%	0%	0%	0%
PVC	100%	0%	0%	0%	99%	25%	33%	29%	74%	1%	33%	1%	42%	0%	33%	1%
Ductile Iron	97%	88%	50%	64%	96%	83%	36%	50%	96%	83%	36%	50%	96%	83%	36%	50%
Cluster 0	100%	0%	0%	0%	100%	0%	0%	0%	72%	1%	100%	1%	99%	0%	0%	0%
Cluster 1	93%	92%	65%	76%	93%	90%	67%	77%	92%	88%	65%	75%	94%	89%	74%	81%
Cluster 2	91%	0%	0%	0%	94%	50%	50%	50%	54%	0%	0%	0%	89%	0%	0%	0%
Cluster 3	**Only non-broken pipes**															

FIGURE 0.3 - COMPARING RESULTS FOR DIFFERENT GROUPS (BARRIE – CLASSIFICATION MODELS)

COMPARISON	Random Forest				XGBOOST				Logisitic Regression				ANN				
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	
All Materials	96%	94%	70%	80%	**97%**	93%	81%	87%	**95%**	87%	67%	76%	**97%**	91%	79%	85%	
Cast Iron	92%	94%	80%	86%	94%	93%	86%	89%	86%	74%	82%	78%	92%	90%	83%	86%	
Copper	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	
PVC	**99%**	0%	0%	0%	**99%**	0%	0%	0%	64%	1%	59%	2%	**99%**	0%	0%	0%	
AC	**96%**	**99%**	79%	88%	**96%**	**95%**	80%	86%	**95%**	**97%**	71%	82%	**96%**	**96%**	79%	86%	
Steel	89%	69%	75%	72%	91%	71%	83%	77%	87%	68%	54%	60%	92%	74%	83%	78%	
Cluster 0	90%	90%	90%	90%	92%	91%	94%	92%	89%	89%	90%	90%	89%	92%	87%	89%	**All lined**
Cluster 1	91%	92%	73%	82%	90%	88%	75%	81%	83%	67%	78%	72%	90%	87%	77%	82%	**Unlined**
Cluster 2	**99%**	91%	36%	51%	**99%**	67%	36%	47%	**99%**	33%	18%	23%	**99%**	67%	36%	47%	**Unlined**
Cluster 3	**97%**	**98%**	81%	88%	**96%**	**96%**	80%	87%	**95%**	**97%**	74%	84%	**97%**	**97%**	82%	89%	**Unlined**

FIGURE 0.4 - COMPARING RESULTS FOR DIFFERENT GROUPS (SASKATOON – CLASSIFICATION MODELS)

COMPARISON	Random Forest				XGBOOST				Logisitic Regression				ANN				Note
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	
All Materials	98%	98%	76%	85%	98%	96%	77%	85%	98%	94%	72%	81%	98%	94%	80%	86%	
Cast Iron	94%	100%	90%	95%	94%	100%	90%	95%	96%	100%	93%	96%	95%	93%	95%	94%	
Concrete	**Not enough class one**																
PVC	99%	0%	0%	0%	99%	0%	0%	0%	77%	3%	83%	5%	99%	0%	0%	0%	
Ductile Iron	97%	99%	89%	94%	97%	99%	91%	94%	95%	95%	83%	89%	96%	94%	88%	91%	
Cluster 0	99%	100%	5%	9%	99%	33%	5%	8%	80%	5%	86%	9%	99%	40%	10%	15%	**All lined**
Cluster 1	97%	98%	87%	92%	98%	96%	93%	94%	97%	96%	87%	91%	96%	89%	91%	90%	**DI**
Cluster 2	**Not Enough Samples for brokn pipes**																
Cluster 3	91%	95%	90%	93%	94%	100%	90%	95%	96%	100%	93%	96%	93%	97%	90%	94%	**CI**

FIGURE 0.5 - COMPARING RESULTS FOR DIFFERENT GROUPS (MARKHAM – CLASSIFICATION MODELS)

COMPARISON	Random Forest				XGBOOST				Logisitic Regression				ANN				Note
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	
All Materials	98%	97%	75%	85%	97%	93%	78%	85%	97%	93%	71%	81%	97%	91%	77%	83%	
Cast Iron	91%	98%	81%	89%	91%	94%	84%	89%	89%	89%	83%	86%	90%	92%	84%	88%	
Asbestos Cement	96%	83%	62%	71%	97%	86%	75%	80%	96%	83%	62%	71%	94%	62%	62%	62%	
Concrete	99%	0%	0%	0%	99%	0%	0%	0%	82%	2%	50%	5%	99%	0%	0%	0%	
PVC	99%	0%	0%	0%	99%	0%	0%	0%	72%	2%	46%	3%	99%	0%	0%	0%	
Ductile Iron	94%	94%	78%	85%	94%	92%	78%	84%	93%	84%	84%	84%	93%	90%	79%	84%	
Cluster 0	99%	0%	0%	0%	99%	0%	0%	0%	71%	1%	42%	2%	99%	0%	0%	0%	**Unlined Pipes**
Cluster 1	93%	100%	81%	89%	91%	93%	83%	88%	91%	96%	77%	86%	91%	91%	84%	87%	**Unlined Pipes**
Cluster 2	93%	94%	75%	83%	93%	90%	79%	84%	93%	94%	73%	82%	93%	89%	80%	84%	**Lined Pipes**
Cluster 3	100%	0%	0%	0%	98%	0%	0%	0%	78%	2%	100%	4%	100%	0%	0%	0%	**Unlined Pipes**

FIGURE 0.6 - COMPARING RESULTS FOR DIFFERENT GROUPS (REGION OF DURHAM – CLASSIFICATION MODELS)

COMPARISON	Random Forest				XGBOOST				Logisitic Regression				ANN				Note
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	
All Materials	95%	91%	69%	79%	95%	88%	72%	79%	94%	84%	68%	75%	95%	89%	72%	79%	
Cast Iron	89%	89%	82%	85%	87%	84%	81%	82%	89%	88%	82%	85%	90%	88%	84%	86%	
Concrete	93%	0%	0%	0%	93%	0%	0%	0%	68%	9%	40%	14%	51%	3%	20%	5%	
PVC	95%	0%	0%	0%	95%	0%	0%	0%	60%	6%	50%	11%	93%	0%	0%	0%	
Ductile Iron	97%	0%	0%	0%	97%	47%	20%	28%	79%	8%	62%	13%	97%	17%	5%	8%	
Cluster 0	**Only Copper found in this cluster, only class 0**																
Cluster 1	98%	96%	64%	77%	97%	92%	63%	74%	97%	92%	57%	71%	97%	92%	64%	75%	**Lined pipes**
Cluster 2	99%	0%	0%	0%	99%	0%	0%	0%	85%	8%	100%	14%	97%	0%	0%	0%	
Cluster 3	94%	90%	80%	85%	93%	88%	81%	84%	93%	89%	80%	84%	93%	89%	80%	84%	

FIGURE 0.7 - COMPARING RESULTS FOR DIFFERENT GROUPS (HALIFAX – CLASSIFICATION MODELS)

COMPARISON	Random Forest				XGBOOST				Logisitic Regression				ANN				Note
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	
All Materials	96%	97%	48%	64%	96%	82%	51%	63%	96%	89%	44%	59%	96%	88%	52%	66%	
Cast Iron	91%	85%	71%	77%	91%	84%	73%	78%	89%	77%	70%	73%	90%	81%	74%	77%	
PVC	100%	0%	0%	0%	100%	0%	0%	0%	66%	1%	75%	2%	99%	0%	0%	0%	
Ductile Iron	95%	50%	4%	7%	95%	33%	100%	15%	77%	12%	63%	21%	95%	25%	2%	4%	
Cluster 0	93%	87%	78%	82%	90%	74%	82%	78%	90%	79%	73%	76%	93%	86%	77%	81%	**Unlined Pipes**
Cluster 1	100%	0%	0%	0%	100%	0%	0%	0%	74%	1%	75%	2%	99%	0%	0%	0%	**Unlined Pipes**
Cluster 2	95%	40%	4%	7%	94%	31%	17%	22%	76%	12%	65%	20%	95%	23%	6%	9%	**Unlined Pipes**
Cluster 3	80%	100%	33%	50%	70%	0%	0%	0%	70%	50%	33%	40%	30%	30%	100%	46%	**Lined Pipes**

FIGURE 0.8 - COMPARING RESULTS FOR DIFFERENT GROUPS (KITCHENER – CLASSIFICATION MODELS)

COMPARISON	Random Forest				XGBOOST				Logisitic Regression				ANN				Note
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	
All Materials	93%	88%	37%	52%	94%	82%	52%	64%	78%	26%	72%	38%	93%	72%	41%	52%	
Cast Iron	89%	83%	51%	63%	90%	76%	68%	72%	68%	34%	73%	46%	88%	85%	45%	59%	
PVC	**Not enough broken pipes**																
Ductile Iron	97%	50%	8%	13%	96%	33%	15%	21%	66%	7%	81%	13%	97%	50%	12%	19%	
Cluster 0	97%	0%	0%	0%	97%	17%	4%	7%	70%	7%	75%	13%	97%	33%	4%	7%	**Majority ductile iron**
Cluster 1	87%	86%	41%	55%	89%	78%	63%	70%	71%	37%	71%	49%	88%	87%	43%	57%	
Cluster 2	**All Concrete - Not enough class 1**																
Cluster 3	**All PVC - Not enough class 1**																

FIGURE 0.9 - COMPARING RESULTS FOR DIFFERENT GROUPS (ST. JOHN'S – CLASSIFICATION MODELS)

COMPARISON	Random Forest				XGBOOST				Logisitic Regression				ANN				Note
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	
All Materials	98%	96%	79%	86%	98%	94%	82%	88%	97%	94%	71%	81%	98%	95%	81%	87%	
All Materials - ASR	95%	96%	83%	89%	95%	94%	85%	89%	94%	96%	81%	88%	95%	95%	83%	89%	
Cast Iron	93%	97%	83%	90%	93%	93%	88%	90%	91%	93%	80%	86%	93%	98%	82%	90%	
Ductile Iron	97%	90%	84%	86%	96%	89%	83%	86%	95%	75%	90%	82%	96%	89%	84%	86%	
PVC	100%	0%	0%	0%	100%	0%	0%	0%	70%	1%	93%	1%	95%	50%	7%	12%	
AC	99%	96%	100%	98%	99%	96%	100%	98%	98%	92%	100%	96%	97%	88%	95%	91%	
Concrete	98%	0%	0%	0%	99%	100%	33%	50%	73%	5%	67%	10%	97%	0%	0%	0%	
Steel	99%	67%	50%	57%	99%	40%	50%	44%	86%	4%	75%	8%	98%	20%	50%	29%	
Cluster 0	94%	98%	84%	90%	94%	95%	86%	90%	92%	95%	82%	88%	93%	95%	85%	90%	**Mostly Cast Iron**
Cluster 1	97%	90%	87%	88%	97%	91%	85%	88%	94%	73%	91%	81%	97%	89%	87%	88%	**Mostly Ductile Iron**
Cluster 2	98%	0%	0%	0%	99%	100%	40%	57%	72%	4%	60%	7%	98%	0%	0%	0%	**Mostly Concrete**
Cluster 3	100%	100%	8%	14%	100%	43%	12%	18%	81%	2%	88%	3%	100%	62%	19%	29%	

FIGURE 0.10 - COMPARING RESULTS FOR DIFFERENT GROUPS (CALGARY – CLASSIFICATION MODELS)

COMPARISON	Random Forest				XGBOOST				Logisitic Regression				ANN			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
All Materials	99%	100%	9%	16%	99%	67%	16%	26%	67%	3%	91%	5%	99%	77%	16%	26%
Cast Iron	99%	88%	15%	26%	99%	76%	17%	28%	65%	4%	76%	7%	98%	71%	16%	26%
Ductile Iron	100%	0%	0%	0%	100%	0%	0%	0%	62%	0%	29%	0%	100%	0%	0%	0%
Cluster 0	100%	0%	0%	0%	100%	0%	0%	0%	60%	0%	60%	0%	46%	0%	60%	0%
Cluster 1	98%	86%	6%	12%	98%	45%	10%	16%	65%	3%	72%	6%	98%	45%	11%	17%
Cluster 2	**Only Broken Pipes**															
Cluster 3	**Only Broken Pipes**															

FIGURE 0.11 - COMPARING RESULTS FOR DIFFERENT GROUPS (VANCOUVER – CLASSIFICATION MODELS)

COMPARISON	Random Forest				XGBOOST				Logisitic Regression				ANN				Note
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	
All Materials	95%	87%	31%	46%	95%	81%	42%	55%	94%	71%	24%	36%	95%	78%	38%	51%	
Cast Iron	89%	82%	42%	56%	91%	81%	58%	67%	83%	48%	73%	58%	89%	84%	40%	58%	
PVC	**Not enough broken pipes**																
Ductile Iron	94%	67%	13%	22%	93%	50%	20%	29%	77%	16%	60%	26%	94%	67%	27%	38%	
Cluster 0	91%	96%	47%	63%	89%	75%	53%	62%	83%	50%	75%	60%	87%	72%	41%	52%	**Unlined Pipes**
Cluster 1	**Unlined Pipes - Not enough broken pipes**																
Cluster 2	94%	89%	44%	59%	94%	77%	56%	65%	81%	32%	72%	44%	89%	48%	56%	51%	**Lined Pipes**
Cluster 3	95%	80%	27%	40%	95%	100%	27%	42%	73%	17%	73%	27%	95%	83%	33%	48%	**Unlined Pipes**

FIGURE 0.12 - COMPARING RESULTS FOR DIFFERENT GROUPS (WATERLOO – CLASSIFICATION MODELS)

COMPARISON	Random Forest				XGBOOST				Logisitic Regression				ANN			
	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
All Materials	96%	91%	58%	70%	96%	86%	66%	75%	94%	73%	43%	54%	96%	85%	65%	73%
Cast Iron	87%	87%	63%	73%	87%	83%	69%	75%	77%	59%	73%	65%	86%	82%	67%	74%
PVC	100%	100%	2%	4%	100%	14%	2%	3%	72%	1%	75%	2%	75%	0%	16%	0%
Ductile Iron	93%	87%	87%	87%	92%	86%	84%	85%	84%	65%	82%	72%	92%	85%	85%	85%
AC	100%	100%	2%	4%	100%	14%	2%	3%	72%	1%	75%	2%	75%	0%	16%	0%
Copper	**Uncoated - Not enough broken pipes**															
Cluster 0	100%	95%	53%	68%	100%	91%	51%	65%	99%	78%	38%	51%	100%	92%	54%	68%
Cluster 1	96%	95%	64%	77%	96%	93%	66%	77%	95%	92%	56%	70%	96%	91%	64%	75%
Cluster 2	**StyroFoam - only class 0**															
Cluster 3	86%	85%	64%	73%	87%	84%	70%	77%	77%	60%	72%	66%	86%	83%	68%	75%

FIGURE 0.13 - COMPARING RESULTS FOR DIFFERENT GROUPS (WINNIPEG – CLASSIFICATION MODELS)

APPENDIX I – PYTHON CODES FOR CLASSIFICATION MODELS

```
## Import tools

import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import statistics
from scipy import stats
from sklearn.metrics import plot_confusion_matrix, classification_report,
plot_roc_curve, plot_precision_recall_curve, accuracy_score, recall_score,
precision_score, roc_auc_score, f1_score, auc


# define Data Frame

df = pd.read_csv("../Data/Cleaned_Classification_New.csv")


# plot correlation matrix

corr_matrix = df.corr(method = 'spearman')
fig, ax = plt.subplots(figsize = (8,5), dpi = 200)
ax = sns.heatmap(corr_matrix,annot = True, linewidths = 0.5, fmt =".2f",cmap
= "YlGnBu")


# plot the number of each class

plt.figure(figsize = (6,4), dpi = 200)
sns.countplot(data = df, x = "Target"), print(df["Target"].value_counts())


# change categorical attributes to dummy variables

df_ = pd.get_dummies(df)

corr_df = pd.get_dummies(df).corr()


# plot correlation matrix including all categorical attributes

corr_matrix = df_.corr(method = 'spearman')
fig, ax = plt.subplots(figsize = (25,15), dpi = 300)
ax = sns.heatmap(corr_matrix,annot = True, linewidths = 0.5, fmt =".2f",cmap
= "YlGnBu")


# define X and y for prdiction
```

```
X = df_.drop("Target", axis = 1)
y = df_["Target"]
```

Random Forest RandomizedSearchCV

```
from sklearn.ensemble import RandomForestClassifier

## Create Splits for RandomizedSearchCV

from sklearn.model_selection import train_test_split

X_train,X_test,y_train,y_test = train_test_split(X,y,test_size = 0.2,
random_state=101)

base_rf_model = RandomForestClassifier()


# define a range of parameteres for Random Forest model to be used in
RandomizedSearchCV

param_grid =
{"n_estimators":np.arange(64,300,4),"max_depth":[5,6,7,8,9,10],"criterion":["
gini","entropy"],"warm_start":[True,False],"max_features":["auto","sqrt","log
2"]}



from sklearn.model_selection import RandomizedSearchCV

rf_grid_model =
RandomizedSearchCV(base_rf_model,param_distributions=param_grid,cv=5,verbose=
3, scoring = "f1",n_iter=20)

## Fit model

%%time
rf_grid_model.fit(X_train,y_train)

## Extract best parameters found by Random Forest

rf_grid_model.best_params_

# make prediction

rf_pred_grid = rf_grid_model.predict(X_test)

# Classification Reprt

print(classification_report(y_test,rf_pred_grid))
```

```
# Plot Confusion Matrix

plt.figure(figsize = (20,8))
plot_confusion_matrix(rf_grid_model,X_test,y_test);

# Defince different metrics for evaluation

accuracy_rf_grid = accuracy_score(y_test,rf_pred_grid)
f1_rf_grid = f1_score(y_test,rf_pred_grid)
precision_rf_grid = precision_score(y_test,rf_pred_grid)
recall_rf_grid = recall_score(y_test,rf_pred_grid)
auc_rf_grid = roc_auc_score(y_test, rf_grid_model.predict_proba(X_test)[:,
1])

# Plot 1st Tree of Random Forest


rf_tuned = RandomForestClassifier(n_estimators=148,max_depth=7,
criterion="gini",max_features="sqrt",warm_start=True)
rf_tuned.fit(X_train,y_train)
from sklearn.tree import plot_tree
from sklearn import tree
plt.figure(figsize=(80,15), dpi = 150)
tree.plot_tree(rf_tuned.estimators_[1],feature_names=X.columns, filled =
True,fontsize=7, class_names= True);
```

Visualize the most important features (Random Forest)

```
rf_tuned.feature_importances_
imp_feats_rf_grid = pd.DataFrame(data=rf_tuned.feature_importances_,
                index = X.columns,columns = ["Feat Imp"])
imp_feats_rf_grid = imp_feats_rf_grid.sort_values("Feat Imp", ascending =
False )
imp_feats_rf_grid

# print out the most important features as a bar plot

plt.figure(figsize = (10,3), dpi = 200)
sns.barplot(data = imp_feats_rf_grid, x = imp_feats_rf_grid.index, y = "Feat
Imp")
plt.xticks(rotation=90);
```

APPENDIX J – PYTHON CODES FOR REGRESSION MODELS

```
## Import tools

import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import statistics
from scipy import stats
# define Data Frame

df = pd.read_csv("../Data/Regression-ROF-New.csv")

# plot correlation matrix

corr_matrix = df.corr(method = 'spearman')
fig, ax = plt.subplots(figsize = (10,5), dpi = 200)
ax = sns.heatmap(corr_matrix,annot = True, linewidths = 0.5, fmt =".2f",cmap
= "YlGnBu")


# change categorical attributes to dummy variables

df_ = pd.get_dummies(df)

corr_df = pd.get_dummies(df).corr()


# define X and y for prdiction

X = df_.drop("CurrentRoF", axis = 1)
y = df_["CurrentRoF"]
```

Artificial Neural Networks (ANN) – Multi Layer Perceptron Regressor (MLPRegressor)

```
from sklearn.neural_network import MLPRegressor


## Create Splits for RandomizedSearchCV

np.random.seed(60)
from sklearn.model_selection import train_test_split

X_train,X_test,y_train,y_test = train_test_split(X,y,test_size = 0.25,
random_state=101)
```

```
# scale dataset into uniform range of values
# we do fit scalar only on X_train

np.random.seed(60)
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
scaled_X_train = scaler.fit_transform(X_train)
scaled_X_test = scaler.transform(X_test)


base_NN_model = MLPRegressor()


np.random.seed(60)
# define a range of parameteres for MLPRegressor model to be used in
RandomizedSearchCV

param_grid =
{"hidden_layer_sizes":[(22,),(22,20,),(22,20,20,18,14,),(22,21,15,8,)],"activ
ation":["relu","logistic","tanh"],"solver":["adam","sgd","lbfgs"],"alpha":[0.
001,0.0001,0.002,0.02,0.00001,0.005],"learning_rate":["adaptive","invscaling"
,"constant"],"max_iter":[2500,5000],"warm_start":[True,False],"momentum":[0.1
,0.001,0.002,0.005,0.3,0.6,0.9],"early_stopping":[True]}


from sklearn.model_selection import RandomizedSearchCV

NN_grid_model =
RandomizedSearchCV(base_NN_model,param_distributions=param_grid,cv=5,verbose=
3, scoring= "neg_mean_squared_error",n_iter=60)


# Fit Model

np.random.seed(60)
NN_grid_model.fit(scaled_X_train,y_train)


NN_grid_model.best_estimator_


# make prediction
NN_pred_grid = NN_grid_model.predict(scaled_X_test)


from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score


MAE_NN = mean_absolute_error(y_test,NN_pred_grid)
MAE_NN


MSE_NN = mean_squared_error(y_test,NN_pred_grid)
```

```
MSE_NN


r2_NN = r2_score(y_test,NN_pred_grid)
r2_NN


RMSE_NN = np.sqrt(MSE_NN)
RMSE_NN


print({"MAE": MAE_NN,"MSE":MSE_NN,"R2":r2_NN,"RMSE":RMSE_NN})


# plot adaboost regression graoh for Age to first Failure
plt.figure(figsize=(6,4),dpi=200)
sns.regplot(x = y_test,y = NN_pred_grid)
```

APPENDIX K – HYPER PARAMETERS (CLASSIFICATION MODELS)

TABLE 0.1~~0.1~~ – RANDOM FOREST HYPERPARAMETERS (ALL MATERIALS)

Utility	Random Forest Hyper parameters (All materials Category)			
	n_estimators	max_depth	max_features	criterion
Saskatoon	148	10	sqrt	gini
Winnipeg	236	10	auto	gini
Kitchener	272	10	log2	gini
Markham	250	9	auto	gini
Waterloo	100	9	auto	gini
Region of Waterloo	100	8	auto	gini
Region of Durham	110	9	auto	gini
Calgary	100	9	sqrt	entropy
Vancouver	100	20	log2	entropy
Victoria	100	9	sqrt	entropy
Halifax	100	9	auto	entropy
St. John`s	100	9	auto	gini
Barrie	100	9	auto	gini

TABLE 0.2~~0.2~~ - XGBOOST HYPERPARAMETERS (ALL MATERIALS)

Utility	XGBOOST Hyper parameters (All materials Category)					
	sampling method	min_child_weight	lambda	gamma	eta	alpha
Saskatoon	gradient_based	0.42	0.895	0.267	0.371	0.216
Winnipeg	gradient_based	0.397	0.166	0.319	0.216	0.035
Kitchener	uniform	0.1	0.5	0.1	0.3	0.1
Markham	uniform	0.25	1	0.3	0.3	0.25
Waterloo	gradient_based	0.25	0.5	0.3	0.3	0.25
Region of Waterloo	uniform	0.25	0.5	0.05	0.5	0.25
Region of Durham	gradient_based	0.25	1	0.1	0.3	0.1
Calgary	gradient_based	0.25	0.5	0.3	0.3	0.25
Vancouver	gradient_based	0.25	0.25	0.1	0.3	0.1
Victoria	uniform	0.5	1	0.3	0.3	0.25
Halifax	gradient_based	0.25	0.5	0.1	0.3	0.1
St. John`s	gradient_based	0.1	1	0.3	0.5	0.1
Barrie	uniform	0.25	0.25	0.3	0.3	0.1

TABLE 0.3~~0.3~~ – LOGISTIC REGRESSION HYPERPARAMETERS (ALL MATERIALS)

Utility	Logistic Regression Hyper parameters (All materials Category)			
	solver	penalty	class_weight	C
Saskatoon	lbfgs	none	dict	0.448
Winnipeg	lbfgs	none	dict	0.653
Kitchener	newton-cg	none	dict	1
Markham	newton-cg	none	dict	1
Waterloo	newton-cg	none	dict	1
Region of Waterloo	sag	none	balanced	1
Region of Durham	newton-cg	none	dict	1
Calgary	newton-cg	none	dict	1
Vancouver	newton-cg	l2	balanced	1
Victoria	newton-cg	none	dict	1
Halifax	saga	l1	dict	1
St. John`s	newton-cg	l2	balanced	1
Barrie	newton-cg	l2	dict	1

TABLE 0.4~~0.4~~ – ARTIFICIAL NEURAL NETWORKS (ANN) HYPERPARAMETERS (ALL MATERIALS)

Utility	ANN Hyper parameters (All materials Category)						
	solver	momentum	max_iter	learning_rate	hidden layer	alpha	activation function
Saskatoon	lbfgs	0.3	2500	adaptive	(26,)	0.001	logistic
Winnipeg	lbfgs	0.3	2500	constant	(26,)	0.0001	relu
Kitchener	lbfgs	0.3	2500	constant	(19,)	0.0001	relu
Markham	lbfgs	0.3	2500	invscaling	(19,)	0.0001	relu
Waterloo	lbfgs	0.001	5000	invscaling	(18,)	0.002	relu
Region of Waterloo	lbfgs	0.9	2500	adaptive	(16,14,)	0.02	relu
Region of Durham	lbfgs	0.3	2500	constant	(18,)	0.005	relu
Calgary	lbfgs	0.9	2500	adaptive	(14,12,)	0.002	relu
Vancouver	lbfgs	0.9	2500	adaptive	(21,19,)	0.02	relu
Victoria	lbfgs	0.005	2500	adaptive	(20,)	0.005	relu
Halifax	lbfgs	0.005	2500	adaptive	(20,)	0.005	relu
St. John`s	lbfgs	0.001	5000	invscaling	(14,14,)	0.005	relu
Barrie	lbfgs	0.3	2500	invscaling	(23,)	0.0001	relu