EVALUATION OF DECISION-MAKING PREDICTION MODELS FOR SEWER PIPES ASSET MANAGEMENT

by

SALAR SHIRKHANLOO

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

August 2022

Dedication

This dissertation is dedicated to my beloved mother, Fatemeh Shirkhanloo, and my dear brothers, Sohrab and Siamak Shirkhanloo, for their endless support, encouragement, patience, and unconditional love, and the memories of my great father, Hamidreza Shirkhanloo, who inspired me to start, continue and complete this journey.

.

Acknowledgements

I would like to express my most sincere gratitude and appreciation to my academic advisor and mentor, Dr. Mohammad Najafi, P.E., F. ASCE, Associate Professor of Civil Engineering and Director of the Center for Underground Infrastructure Research and Education (CUIRE). It was a pleasure working under the supervision of Dr. Najafi. He was always there to help me when I was in need, and I am indebted for the opportunities he provided me.

I should express my heartfelt thanks to Dr. Ardeshir Anjomani, Dr. Sharareh Kermanshachi, Dr. Vinayak Kaushal, and Dr. Nilo Tsung for their continued support and guidance as my dissertation committee members. Their valuable comments and suggestions made a great contribution to this dissertation.

I would like to extend my sincere thanks to Dr. Mohammadreza Malek Mohammadi, Engineer/Data Scientist at Plummer Associates, Inc., Dr. Daniel Atambo, Senior Engineer/Project Manager at the City of Dallas, Dallas Water Utilities, and Dr. Karthikeyan Loganathan, Assistant Professor of Instruction at the University of Texas at Arlington for their continued commitment to providing all the data and information required for this research study. I would like to extend my special thanks to Dr. Aireza Fallahi, Data Scientist at American Airlines, and Mrs. Madhuri Arjun, Graduate Teaching Assistant at the University of Texas Arlington, for their keen interest in my research study and endless support throughout the model development phases. Without their help, I wouldn't have accomplished this study.

It would be ungrateful if I forgot to thank my lovely deceased father, my beloved mother, and my dear brothers. They have always been the most important source of motivation during my PhD studies and in other aspects of my life.

July 31, 2022

Abstract

EVALUATION OF DECISION-MAKING PREDICTION MODELS

FOR SEWER PIPES ASSET MANAGEMENT

Salar Shirkhanloo, PhD

The University of Texas at Arlington, 2022

Supervising Professor: Dr. Mohammad Najafi

Wastewater collection systems deteriorate over time, requiring continuous adjustments and the development of asset management frameworks on the part of utility owners to maintain the performance of their assets. Any asset management framework should emphasize the importance of asset inspection and condition evaluation for efficient system operation and maintenance. Closed-circuit television (CCTV) is the most widely used tool in the United States for inspecting the interior of sewer pipes, which is a somewhat expensive and time-consuming process given the extensive inventory of pipes in a city. Due to their vast inventory of these pipes, no municipality can inspect every individual sanitary sewer pipe section in a short amount of time. Therefore, the main goal of this research is to develop prediction models that can anticipate the future state of sewer pipes. The results of the models can be used to rank the necessity for sanitary sewer pipe inspection, rehabilitation, and replacement. Combined data collected from the City of Dallas, Texas, and the City of Tampa, Florida, were used in this dissertation. This dataset included nine independent variables: pipe age, size, length, material, surrounded soil type, soil pH, depth, slope, and surface conditions, and one dependent variable was the condition rating of sewer pipe based on PACP scores from 1 to 5. Different resampling procedures were examined in this study to overcome the problem of the imbalanced dataset, and finally, the resampled dataset by the SVM-SMOTE method was selected. Various machine learning algorithms such as Logistic Regressions, k-nearest neighbors, Decision trees, Random Forests, AdaBoost, Gradient Boosting Tree, and XGBoost were employed to develop prediction

models. The other objective of this dissertation is an investigation of the efficiency of different machine learning methods using a resampled dataset, which was done thoroughly in this study. Various evaluation metrics, including precision, recall, F1-score (see Section 3.5.5), and area under the curve (AUC), were calculated to compare the effectiveness of developed models. The overall F1-score for the Random Forest model was 0.80 and for Multinomial Logistic Regression was 0.48, which were the highest and lowest, respectively. It was concluded that tree-based models had better performance than other models and the bagging approach was more efficient than boosting. Additionally, as another objective of this dissertation, using the best model results, it was found that pipe age and length had the highest effect on the condition rate of sewer pipes, while pipe location had the least impact.

NOTE: Please refer to Appendix A for abbreviations.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction and Background

1.1 Introduction

The underground infrastructure networks cover thousands of kilometers and are an important part of the overall infrastructure of the United States (Najafi and Gokhale, 2022). Sanitary sewers are intended to collect sanitary sewage from residential, industrial, commercial, and public users and convey it to a treatment plant as part of wastewater infrastructure systems. The majority of sewer systems are gravity sewers, which transmit flow based on an initial slope. Over 240 million Americans are associated with 14,748 wastewater treatment plants, with an estimated 56 million people expected to use centralized treatment plants by 2032 (ASCE, 2020).

Some elements of the wastewater infrastructure in the United States are over a century old, and a combination of age, malfunctions, and accidents results in at least 23,000 to 75,000 sanitary sewer overflows per year (EPA, 2004). The American Society of Civil Engineering (ASCE) has released its 2020 Infrastructure Report Card, which gives wastewater infrastructure a "D plus" score. According to ASCE, water and wastewater infrastructure in the United States are plainly aging, and a $150 billion capital budget deficit is expected by 2025 to keep up with the demand (ASCE, 2020).

Sewer pipes are an important part of wastewater systems because they connect wastewater generating points to treatment plants. The structural and operational performance of sewer systems deteriorates as they age. Sewer pipe aging increases failure rates, and deteriorated pipes could have a variety of social, environmental, and economic consequences, including poor water quality, chemical or biological contamination, disease, and high maintenance costs (Opila, 2011).

Rehabilitation techniques are critical factors in maintaining the system's operation at an acceptable standard of support and providing cost-effective alternatives to prevent future failures. Infrastructure asset management is a comprehensive and cost-effective method for keeping pipeline systems in good condition. Asset management programs can establish numerous ways to assist utility companies and municipalities in understanding the timing and related costs of pipeline maintenance, rehabilitation, and replacement (Najafi and Gokhale, 2022).

Sewer pipeline deterioration is a multi-step process that is influenced by a number of variables at the same time. In metropolitan locations, sewer pipes are often hidden and underground, making it difficult to spot pipes with a high risk of collapse. As a result, sewage pipeline inspection and monitoring have received increased attention in recent years in order to avoid additional collapse and failure.

Due to financial, time, and assessment technology constraints, it is evident that monitoring and inspecting all sewer lines is nearly impossible. As a result, greater effort should be put into developing degradation models that can forecast the existing and future state of sewage systems. Several prediction models as well as the variables that impact the state of sewage pipelines will be thoroughly examined in this study.

1.2 Research Needs

Researchers in the United States and throughout the world have conducted hundreds of studies to forecast the state of sanitary sewers. Identifying the important elements influencing sewage pipe degradation and constructing a prediction model based on those factors are two ways to assess the structural integrity of a sanitary sewer pipe. It should be highlighted that the elements are studied for their correlations with the structural integrity of sewage pipes rather than for their causes of breakdowns. As a result, in order for a municipality or utility owner to use a condition prediction model, the model must be able to forecast the state of sanitary sewage pipes using data provided by the municipality or utility owner. Numerous researchers stated that sewer pipe condition prediction models need to be improved, as explained below:

- More historical input variables, such as surface load, groundwater, bedding conditions, soil corrosion, and sewer placement, were recommended by Najafi and Kulandaivel (2005) for improving the neural network model for sewage pipe deterioration.
- For creating sewage pipe condition degradation, Chughtai (2008) recommended using more predictors, such as soil condition, seismic variables, and so on. Future research should also look at the use of different prediction models.
- Salman (2010) suggested that deterioration models be improved by considering additional independent factors such as soil type, groundwater level, and original construction quality.

- According to Opila (2011), further improvement of the condition prediction models would lead to more accurate failure predictions. Other prediction models might be able to deliver more accurate results.
- Sousa et al. (2014) stated that machine learning and artificial intelligence models were more accu rate than logistic regression models and that future research
  might increase findings accuracy.
- According to Kabir et al. (2018), the established sewer structural condition prediction models may be enhanced further by examining the influence of various independent factors as sewer function, groundwater level, soil type, road class, and original construction quality.
- Malek Mohammadi (2019) suggested that instead of changing to binary classes, a prediction model should be able to forecast all five condition levels separately.
- Karthikeyan Loganathan (2021) used various supervised machine learning algorithms to predict sewer pipes conditions and recommended that developed model in his research should be validated on inspection data from a different municipality.

As a result, the knowledge gap in identifying essential components in sewage pipe deterioration has been identified, indicating the necessity for machine learning and artificial intelligence algorithms in the creation of condition prediction models. As a result, one of the study's main aims is to close the knowledge gap found.

1.3 Research Objectives

The first objective of this study is to establish a decision-making support tool as a condition prediction model for sanitary sewer asset management. Assessment of the prediction results of different models for the case study could help cities to design strategic plans for their sewer networks.

The secondary objective of this dissertation is to analyze the differences of the identified significant factors affecting sewer pipe deterioration. Agencies and municipalities can gather fewer data points during inspections by identifying influencing variables.

1.4 Scope of Work

The scope of this dissertation is limited to use of condition scoring system of Pipeline Assessment and Certification Program (PACP) developed by the National Association of Sewer Service Companies (NASSCO). Also, sewer pipes with any rehabilitations are excluded. Table 1-1 shows the scope of this dissertation.

Table 1-1 Scope of the Study

Included	Excluded
Sanitary sewer pipes	Storm-water pipes
Pipes inspected based on PACP manual	Pipes inspected with other manuals
Pipes without any rehabilitation	Pipes with any rehabilitation
PVC, VCP, and RC pipes	Pipes made with other materials

1.5 Research Methodology

Following steps are carried as a methodology to achieve the expected outcome of the research. Figure 1-1 presents more detail about methodology used in this research.

- Step 1: Problem statement
- Step 2: Comprehensive literature review
- Step 3: Data collection
- Step 4: Data analysis for each case study
- Step 5: Development of prediction models for each case study
- Step 6: Model validation for each case study
- Step 7: Comparing artificial intelligence models for each case study
- Step 9: Select the best model for each case study
- Step 10: Compare the results
- Step 11: Conclusions for decision-making

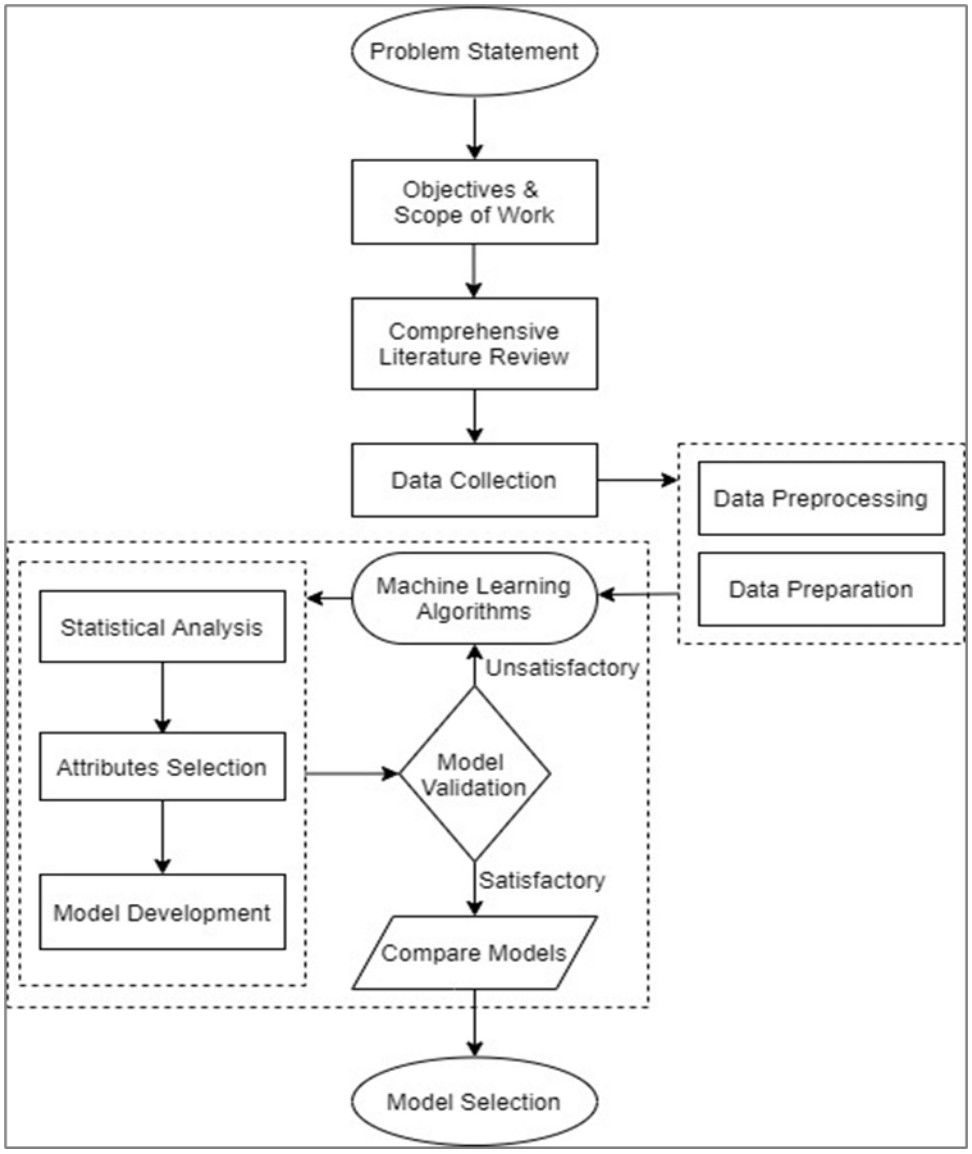

Figure 1-1 Research Methodology

1.6 Expected Outcome

The expected outcomes of this study are:

- A condition prediction model to classify the sewer pipes into multiple classes.
- A comprehensive comparison of the various methods that can be used to select the optimal forecasting model.
- A tool to find the most significant factors affecting the deterioration of pipes.

1.7 Hypothesis

Based on the dataset used in this dissertation, pipe age, material, and length are expected to have the largest effect on sewer failure. In addition, tree-based models are expected to perform better than other statistical and machine learning methods to predict the condition of pipes.

1.8 Chapter Summary

This chapter presented background information on sewage pipe conditions as well as the significance of sewer inspection and maintenance procedures. This chapter also covered the research needs, objectives, scope of work, methodology, expected outcome, and hypotheses.

Chapter 2 Literature Review

2.1 The United States' Sanitary Sewer System

Municipalities began to establish sewage systems to preserve public health and prevent flooding during growing urbanization between 1840 and 1880 (Melosi, 2000). The first comprehensive sewer systems in the United States were built in Chicago and Brooklyn in the late 1850s. Construction of extensive urban sewage systems did not begin until the 1880s. Combined Sewer Systems (CSS) and Separate Sanitary Sewer and Storm Sewer System (SSS) are the two types of sanitary sewer systems used in the United States (EPA, 2004).

In a combined sewer system, a single pipe transports residential, commercial, and industrial wastewater, as well as storm water, to a designated disposal place. During the late nineteenth century, the United States began to create the combined sanitary sewer, taking into account a planned network and big diameter sewers (Burian et al., 2000).

At the turn of the twentieth century, concrete pipes were introduced, followed by polyvinyl chloride, fiberglass, high-density polyethylene, ductile iron, steel, and reinforced concrete pipes (Kulandaivel, 2004). According to the Environmental Protection Agency (EPA) (2010), the following materials are often used in the construction of sanitary and wastewater sewer systems:

- Concrete pipe, including reinforced concrete pipe (RCP) and prestressed concrete cylinder pipe (PCCP).
- Ferrous pipe, including ductile iron, cast iron, and steel.
- Plastic pipe, including polyvinyl chloride (PVC) and high-density polyethylene (HDPE).
- Ceramic-based pipe, including brick and vitrified clay pipe (VCP).

2.2 Asset Management

Asset management in the water and wastewater sector is an adapted idea from several successful applications in other industries such as transportation and building infrastructure management. Asset management was first adopted in Australia and New Zealand in the early 1990s, before spreading to other nations such as Canada, England, and the United States. In the United States, the Federal Highway Administration launched infrastructure asset management in the early 1990s, and the FHWA issued an

Asset Management Primer in 1999. Asset management began to be applied in the water and wastewater industries in the early 2000s, and the Environmental Protection Agency (EPA) played a significant role in providing and supporting asset management principles (Syachrani, 2010).

For wastewater management utilities, asset management can be defined as an inclusive plan to manage infrastructure capital assets to minimize the total cost of owning and operating them, while delivering a satisfactory level of service (EPA, 2004). The main components of Infrastructure management system framework are presented in Figure 2-1.

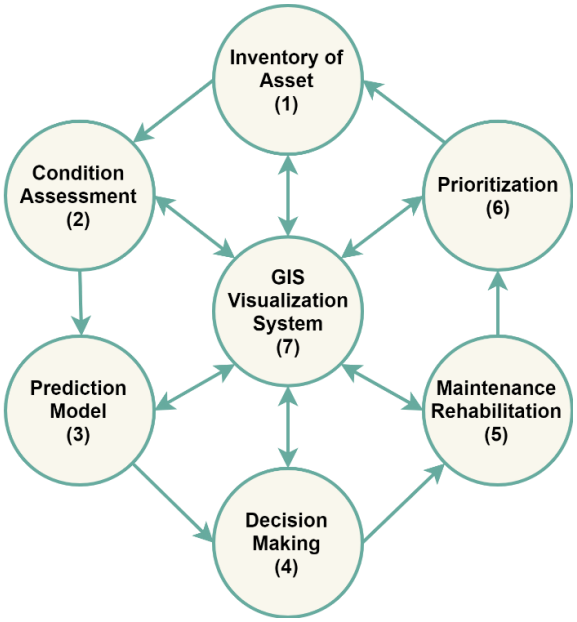

Figure 2-1 Infrastructure Management System Framework

(Malek Mohammadi, 2019)

First step is asset inventory which is process of data collection. The collecting of data is essential in order to execute asset inventory. Utilities or towns in the water and wastewater industry must keep a comprehensive record of assets, including age, location, material, depth, length, and other critical information. Following data collection, information must be analyzed and ranked using condition coding systems. The following step is to develop prediction models to forecast the asset's future status. During this process, historical data is utilized to forecast the asset's future performance in time to prevent any

unforeseen collapse or failure. Infrastructure systems are vital for everyday operations, and municipalities must anticipate their future condition and remaining usable life. The outcome of condition assessment and prediction models leads agencies to develop a decision-making strategy for the asset's current and future state. Several elements, such as available funds, laws, methods of rehabilitation or replacement, and other essential factors, must be considered during the decision-making process. The next phase in an infrastructure management system is asset maintenance and rehabilitation, which is dependent on the outcome of the decision-making process. Finally, all of the above steps aid the government in prioritizing assets for future investment. In today's asset management approach, all infrastructure management procedures are combined with Geographic Information Systems (GIS).

2.3 Condition Assessment of Sewers

2.3.1 Introduction

In the United States, millions of gallons of human and industrial waste are conveyed into wastewater treatment plan through underground sewer systems every day. This process takes place underground (out of view) and maintaining wastewater collection systems is always one of the critical challenges of governments. As the most municipal sewer systems are at least 60 years old, many communities and utilities are paying more attention to assess the condition of their underground pipes and associated infrastructures (EPA, 2015).

The fundamental idea behind sewer condition assessment is to compare the existing structural and operational state of a sewage pipeline to that of a new or like new pipe. The comparison yields a numerical grade for the asset, which represents the current state of underground sewer systems.

2.3.2 Condition Rating Methods of Sewer Pipes

The Water Research Centre (WRc) in the United Kingdom began a five-year research project in 1977 to adopt a technique to assess the health of sewage pipes based on a generic coding system. The condition rating is intended to evaluate the existing state of sewage lines objectively. The two most prevalent pipe condition categories are structural condition and operational condition (Chughtai and Zayed, 2008). Structural condition evaluates the pipe defects, the physical strength of a pipe and the capability of

the pipe to resist external loads, and operational condition indicates the ability of the pipe to meet its service requirements. The result of structural conditions can be used to determine the necessity of pipe rehabilitation or replacement while the operational condition of a pipe indicates the need for cleaning and maintenance (Malek Mohammadi, 2019).

Various methodologies have been introduced in different nations to score the status of underground sewage pipelines, including WRc (Water Research Center) in the United Kingdom, PACP (Pipeline Assessment and Certification Program) in the United States, NRC (National Research Council Canada) in Canada, and WSAA (Water Service Association of Australia) in Australia (Moteleb, 2010). In the United States most of municipalities and agencies use the PACP methodology to assess the condition of sewer pipes.

2.3.3 PACP Condition Grading Method

Pipeline Assessment and Certification Program (PACP) is the North American Standard for pipeline defect identification and assessment to identify the pipe condition and manage the sewer pipe networks. In 2001, National Association of Sewer Service Companies (NASSCO) developed the PACP in partnership with Water Research Center (WRc) to assess the condition of sewer pipes.

Pipe defects and features can be classified into five categories by NASSCO coding system. The defect classification involves; (1) continuous defects, (2) structural defects, (3) operational and maintenance, (4) construction features, and (5) miscellaneous features coding (EPA, 2015). For each type of defect, the numeric codes are used to rank the severity of the pipe defect and capital letters define the type of defect.

The final condition rating is defined from two major categories which are structural and operation and maintenance (O&M). The below list presents the grades and definitions of grades respectively (NASSCO, 2018):

1 - Minor defect grade

2 - Minor to moderate defect grade

3 - Moderate defect grade

4 - Significant defect grade

5 - Most significant defect grades

PACP assess the condition of pipes on a scale of 1 to 5 based on the result obtained from CCTV inspections and operator judgments. Condition 1 determines the pipe is in excellent condition and condition 5 specifies the pipe has failed or is likely to fail. Pipe with condition rating of 5 needs immediate action for rehabilitation or replacement. Table 2-1 provides the PACP condition rating, from the PACP manual.

Table 2-1 PACP Defect Grades

(NASSCO, 2018)

Condition Grade	Description	Time to Failure
5 Immediate Attention	Defects requiring immediate attention	Pipe has failed or is likely to fail within the next five years
4 Poor	Severe defects that will become Grade 5 defects within the foreseeable future	Pipe will probably fail in 5- 10 years
3 Fair	Moderate defects that will continue to deteriorate	Pipe may fail in 10-20 years
2 Good	Defects that have not begun to deteriorate	Pipe unlikely to fail for at least 20 years
1 Excellent	Minor defects	Failure unlikely in the foreseeable future

The outcome of the PACP condition grading method is entirely dependent on the accuracy of the defect coding, and any error during defect identification influences the final grade result. The PACP

condition grading method assigns a ranking to pipe segments depending on the severity of the identified defects and problems.

2.4 Factors Affecting Condition of Sewer Pipe

2.4.1 Introduction

Several attempts have been made in recent years to assess the state of sewage pipes and to identify the factors that impact degradation. Table 2-2 shows these factors. Identification of affecting factors is crucial, according to Kley and Caradot (2013), for the following reasons:

- Data collection is a very expensive process during condition assessment and gathering all the pipe information is not a cost-effective approach. Identification of significant factors decreases the number of required features and reduces data collection costs.
- When more relevant factors are utilized in the model, high prediction accuracy may be attained.

Table 2-2 Factors Affecting Wastewater Pipes Deterioration

(Davies et al., 2001; Al Barqawi and Zayed, 2006)

Physical factors	Environmental factors	Operational factors
Sewer age	Bedding material	Flow velocity
Sewer size	Soil type	Infiltration/exfiltration
Sewer depth	Backfill type	Previous maintenance
Installation method	Surface type	Sediment level
Sewer pipe material	Road type	Surcharge
Joint type	Traffic characteristics	Burst history
Pipe length	Ground movement	Debris
Connections	Groundwater level	Hydraulic condition
Pipe slope	Root interference	Blockages
Pipe shape	Soil corrosivity	Operating pressure (for sewer mains)
Start invert elevation	pH	Sewer function
End invert elevation	Soil fracture potential	
Rim elevation	Vehicle flow	
	Bus flow	
	Number of trees	
	Soil moisture	
	Sulfate soil	

2.4.2 Pipe Age

The difference between the pipe installation year and the date of inspection is commonly referred to as pipe age. The age of a pipe begins the moment it is installed (Kulandaivel, 2004). Various studies have shown that the age of sewer pipes has a significant impact on their condition (Ariaratnam et al. 2001, Kienow and Kienow 2004, Chughtai and Zayed 2008, Ana et al. 2009, Salman and Salem 2012, Laakso et al. 2018). As illustrated in Figure 2-2, the serviceability of pipes deteriorates with time and is separated into five stages (Misiunas 2005).

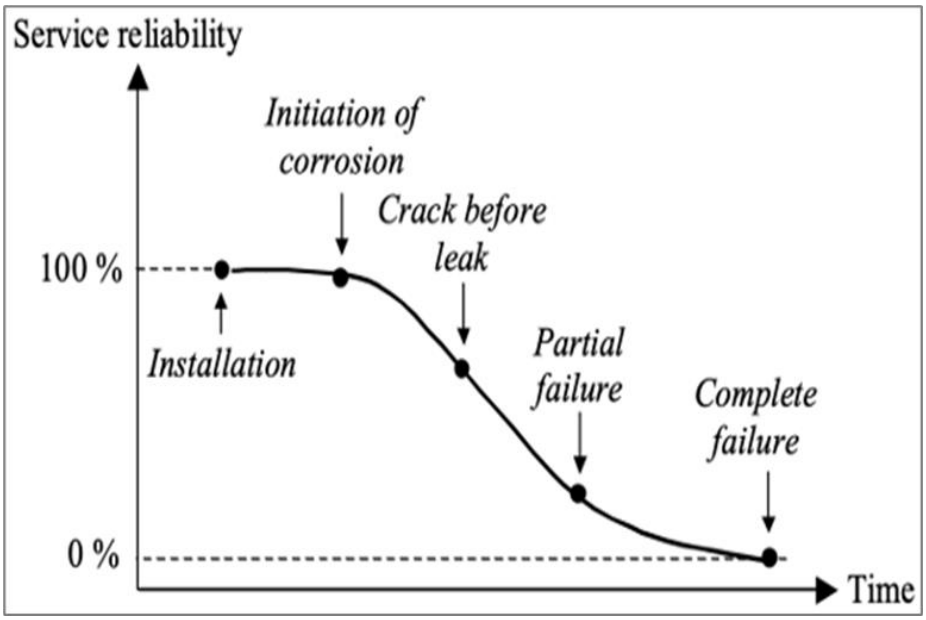

Figure 2-2 Serviceability of a Pipe

(Misiunas, 2005)

Pipe failure is represented in the shape of a bathtub curve, according to Singh and Adachi (2013), which is generated when the pipe failure rate is plotted against timeThe bathtub curve comprises three s eparate stages, as shown in Figure 2-3. The first is the early life period, which has a high failure rate and displays problems soon after installation. Human factors, pipe damage during construction and installation, and improper pipe material can all cause to failures during this period. The second phase indicates the pipe's useful lifetime, with the frequency of failure rate being very low and almost constant. Failures in the second phase might arise in a variety of unpredictable events, such as exceptionally heavy loading, earth movement, settlement, or third-party interference. Finally, due to pipe degradation and age, the third phase (wear-out life) has a significant failure rate (Singh and Adachi, 2013). On the other hand, few studies stated that the age is not a major factor in pipe deterioration (Tafuri and Dzuray, 2000; and Davies et al. 2001).

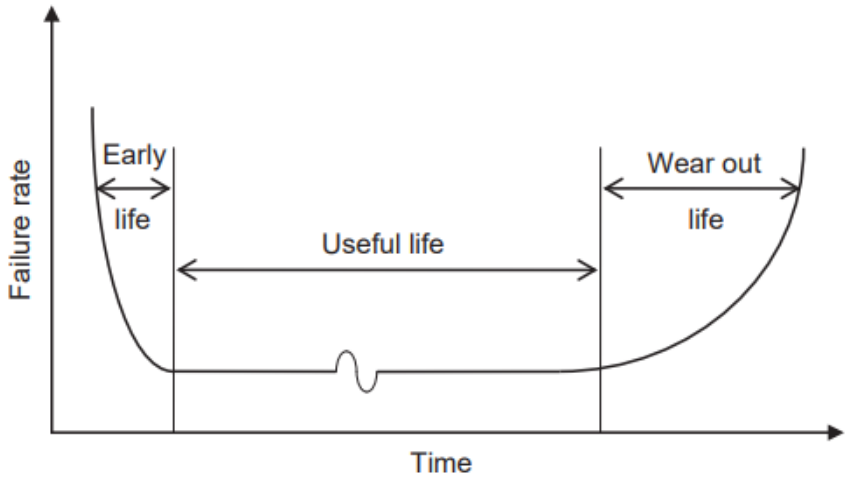

Figure 2-3 The Theoretical Bathtub Curve of Buried Pipe

(Singh and Adachi, 2013)

2.4.3 Pipe Material

Vitrified clay pipe (VCP), ductile iron (DI), cast iron (CI), polyvinyl chloride (PVC), reinforced concrete pipe (RCP), and other materials are used to make sewer pipelines. Each material has its own properties and failure mechanisms would differ as well (Loganathan, 2021). Concrete pipes, for instance, are abrasion-resistant, whereas clay pipes are acid-resistant. Plastic pipes, such as PVC or HDPE, are resistant to acidic and alkaline wastes, although they can deform excessively when loaded (Singh and Adachi, 2013). Concrete pipes behaved better in the model than bricks and clay pipes, according to Ana et al. (2009). Lubini and Fuamba's (2011) model included pipe material as well. They discovered that reinforced concrete pipes are more resistant to degradation than ordinary pipes because the conduit is strengthened with steel, making it robust enough to avoid structural deterioration. According to Laakso et al. (2018) concert and polyethylene high-density pipes were shown to be relevant in the prediction model. Several studies, including Malek Mohammadi (2019), stated that pipe material is a critical factor in pipe deterioration and is among significant parameters in developing prediction models.

2.4.4 Pipe Diameter

Several studies have shown that pipe diameter has an important effect on pipe's deterioration. When the diameter of a sewer pipe is between 6 and 8 inches, it is categorized as a small sewer pipe, and

when the diameter is greater than 10 inches, it is defined as a large sewer pipe (Loganathan, 2021). Some condition prediction models indicated that the rate of sewer pipe deterioration reduces as pipe diameter increases, whereas some other studies have found that smaller diameter pipes fail more frequently (Malek Mohammadi, 2019). Ariaratnam et al. (2001) indicated that when pipe diameter increases the likelihood of a pipe being in a deficient condition decrease. Larger diameter pipes perform better than smaller diameter pipes, according to Lubini and Fuamba (2011), Salman and Salem (2012), and Bakry et al. (2016). Because bigger diameter pipes may continue to run when obstructions are encountered, but smaller diameter pipes lose hydraulic flow. Other studies also have found that because larger pipes are buried deeper, which may be the reason for their better structural condition, they have lower deterioration rates than smaller diameter pipes (Malek Mohammadi et al., 2020: Najafi and Gokhale, 2022).

In contrast, Jeong et al. (2005) stated that larger pipes are more likely to deteriorate, since they have more surface area exposed to sewage and surrounding soil areas. Laakso et al. (2018) found a relatively dual behavior in the variation of pipe diameter and condition levels of sewer pipes. He concluded that pipes with a diameter of 12 and 60 inches were in better condition due to further careful installation supervision. On the other hand, some investigations such as Ana et al. (2009) stated that pipe diameter is not a significant variable in deterioration model.

2.4.5 Pipe Length

The length of a sewage pipe is defined as the distance between the entrance and exit manholes. According to Najafi and Gokhale (2022), shorter pipes deteriorate at a faster rate than longer pipes because longer pipes would have some sharp bends over their length, potentially resulting in less debris or obstructions. On the other hand, Malek Mohammadi (2019) indicated that long sewer pipes could have a greater degradation rate since their probability of flaw is higher.

Also, some studies show a dual behavior in the condition of pipes regarding changes in pipe length. For instance, according to Laakso et al. (2018), sewer pipes longer than 131 feet decay more quickly than other pipes in the network, while pipes shorter than 131 feet have almost no effect on the pipe's condition. Longer pipes have a larger risk of flaws and bending stress, which can explain this consequence. Furthermore, lateral connections can cause structural damage, and longer pipelines have more of them.

2.4.6 Pipe Slope

According to Tran et al. (2006), pipes with steeper slopes are more likely to be vulnerable due to voids in the soil, soil movement, and pipe joint faults. As per Salman and Salem's (2012) prediction model, steeper pipes are more prone to degrade because of stability problems and high flow rates.

In contrast, Laakso et al. (2018) found that extremely low slope was the most dangerous situation for sewer pipes. Extremely low slopes result in insufficient rinsing, resulting in debris accumulation and obstructions. On the other hand, according to the findings of Sousa et al. (2014) and Kabir et al. (2018), pipe slope is a non-significant factor.

2.4.7 Pipe Depth

A sewage pipe's depth is defined as the distance between the pipe's crown and the ground level (Loganathan, 2021). To determine the proper depth of sewage pipes, several elements must be addressed, including soil type, water table, pipe material, pipe diameter, and regulations (Malek Mohammadi, 2019).

Pipe depth is an important variable in Khan et al. (2010)'s prediction model, and each increase in depth has a negative influence on sewer pipe condition level. The higher dead load over the pipes might be the explanation for this behavior.

In contrast, according to some studies, sewer pipes placed at shallow depths deteriorate faster than those buried at deeper depths (Harvey and McBean, 2014; Gedam et al., 2016). In general, increasing the cover thickness above the pipes reduces the impact of surface elements like traffic and construction. Nevertheless, Tran et al. (2006) and Ana et al. (2009) claimed that sewer pipe depth is unimportant in condition prediction modeling.

2.4.8 Sewer Location

The applied load from the surface might have an impact on a sewage pipe. The amount of surface loading carried to the sewage pipe is affected by land use and traffic above the pipe. Surface loads vary in magnitude and frequency, making it difficult to quantify or predict their size (Kley and Caradot, 2013). According to statistics provided by the Federal Highway Administration (FHWA, 2011), highways often have more VMT (Vehicles Miles Traveled), which indicates a more pressure on the surface.

Only a few studies have looked at the impact of pipe location on sewage pipe degradation. Pipe segments beneath local streets and alleyways are less likely to deteriorate than pipe segments beneath gardens or any other form of roadway, according to Salman and Salem (2012). According to Bakry et al. (2016), sewage pipes deteriorate more quickly in industrial zones than in residential areas. In contrast, Micevski et al. (2002) and Tran et al. (2006) stated that the location of the pipe is not an important factor in their prediction model.

2.4.9 Soil Type

Different types of soil have different reactions with pipe material, groundwater, and other pipe attributes or environmental factors (Kaushal and Guleria, 2015). The underlying soil has a considerable influence on sewage pipe deterioration, according to Wirahadikusumah et al. (2001). When comparing pipes placed in stable soil to pipes installed in unstable soil, it was discovered that pipes located in unstable soil suffered more fluctuations in condition (Tafuri and Dzuray 2000). Furthermore, the type of soil around the sewage pipe is one of the most critical aspects that can impact frost heave, soil-pipe interaction strength, and external corrosion, all of which can contribute to failure mechanisms (Najafi and Gokhale, 2022). When there is not enough soil support around a sewage pipe, it might move, causing voids to form around the pipe, making it more prone to deform (Loganathan, 2021). However, in the prediction model developed by Laakso et al. (2018), soil type was not a significant parameter.

2.4.10 Corrosivity

Soil corrosivity is a soil property that enhances the likelihood of external corrosion on pipe surfaces (Malek Mohammadi, 2019). Corrosion in steel pipes is often generated by an electrochemical interaction between the exposed pipe's outer surface and the soil environment surrounding it. Different pipe materials have various degrees of corrosion resistance. The corrosion rate is observed to be impacted by a wide range of variables such as soil acidity, resistivity, pH content, oxidation-reduction, sulfide, moisture, aeration, and so forth (Loganathan, 2021). According to Najafi and Gokhale (2022), longitudinal failure can occur when the pipe wall deteriorates due to corrosion. It should be noted that Only a few research have looked at the influence of soil corrosivity on sewage pipeline deterioration.

2.4.11 Soil pH

The pH of the soil affects the corrosion rate of buried pipelines, according to almost all studies in the field of underground corrosion (Wasim et al., 2018). According to Najafi and Gokhale (2022), soil pH is a good indicator of external corrosion since various pH ranges cause different corrosion processes. Alkaline (pH>7), natural (pH=7), and acidic (pH<7) are the different pH ranges.

Hou et al. (2016) investigated the impact of soil pH on pipes made of various materials. according to the findings, in the same corrosive conditions, cast iron pipes are more likely to corrode than steel pipes. In contrast, some studies stated no relationship between pH and corrosion rate, such as Wasim et al. (2018).

2.4.12 Groundwater Level

Groundwater availability above sewer pipes may result in water running into the pipe (infiltration) through cracks and the loss of soil support (Davies et al., 2001). This infiltration might result in overflows and soil sediments inside the sewage pipes, leading to practical problems. Also, structural troubles are caused by a lack of sufficient support surrounding the pipe. Sewage pipes in places where the groundwater level is much higher are more likely to fail than sewer pipes in areas where the groundwater level is below sewer level, according to Malek Mohammadi et al. (2019). This is due to an increase in the amount of pressure on pipes from groundwater. It should be noted that groundwater levels are often not included in pipeline inventories, and it has been only utilized as a variable in a few prediction models.

2.5 Prediction Models for Sewers

2.5.1 Introduction

Condition prediction models may be used to estimate the condition rating of an infrastructure using data from inspection databases. In general, utility companies and municipalities can estimate the future health of their assets by using degradation models to identify the pipes that need repair, rehabilitation, or replacement. The ultimate objective of many prediction models is to use an acceptable mathematical approach to anticipate the target variables with the maximum accuracy possible.

Deterioration models for sewer pipelines are classified into different categories. Thus, existing sewer deterioration models can be classified into two groups of statistical and artificial intelligence (AI)

models. Statistical models include linear and logistic regressions. AI models includes various machines as shown in Figure 2-4 (Liu et al., 2022).

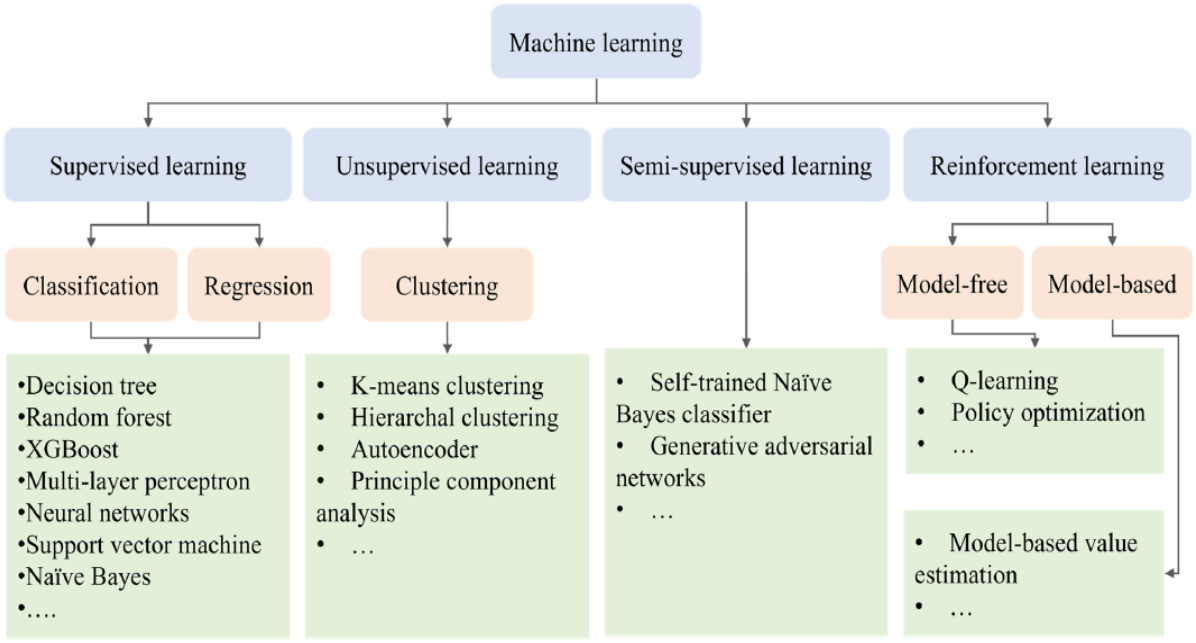

Figure 2-4 Classification of Prediction Models

(Liu et al., 2022)

2.5.2 Statistical Models

A statistical model is defined as a random variable X that represents a quantity whose outcome is uncertain. The probabilistic character of historical data is employed in statistical models to explain model output as a random variable. Estimates are "best guesses" depending on the state of the historical data in any statistical analysis (Coles, 2001). The parametric density function, according to Dasu and Johnson (2003), is used in statistical models to evaluate errors and find probabilistic correlations between dependent and independent variables. Statistical models' outputs and outcomes may be provided in probability values, making them more appropriate to predicting the existing and future state of sewage pipes than deterministic models, which provide quantitative results (Tran, 2007). Numerous statistical models, such as, logistic regression, Markov chain, ordinal regression and cohort survival model were used to predict the condition of sewer pipelines in previous studies.

2.5.2.1 Linear Regression Models

Only one independent variable is used in the simplest linear regression model, and the dependent variable is predicted based on their relationship. As the value of the independent variable rises or falls, the real mean of the dependent variable changes at a constant pace, according to the regression model. As a result, the function connection between the real mean of Yi and Xi, as stated in Eq. 2.1, is represented by the equation of a straight line (Rawlings, 1989).

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad \text{Eq.2.1}$$

Where: i = facility index;

Y_i = dependent variable for facility i;

β_0 and β_1 = parameters to be estimated.

X_i = independent variable.

ϵ_i = random error term

Bakry et al. (2016a, 2016b) developed a condition prediction model for sewage pipes that have previously been restored using the CIPP approach, using regression analysis techniques. Closed-circuit television (CCTV) inspection records of Quebec CIPP rehabilitations were used to compile the information. The models were created based on a number of physical, operational, and environmental aspects. The regression models were evaluated using coefficients of multiple determinations, and the results showed that the accuracy ranged from 80 to 97 percent. Furthermore, mean absolute error and root mean square error were calculated to measure the models' correctness. Generally, the linear regression model is too simple to describe the probabilistic character of pipe degradation and is not suitable for predicting discrete condition values (Tran, 2007; Moteleb, 2010).

2.5.2.2 Logistic Regression

The link between many independent factors and a categorical dependent variable is studied using logistic regressions. The chance of an event occurring is assessed using logistic regression by fitting data to a logistic curve. Binary logistic regression, multinomial logistic regression, and ordinal logistic regression are the three types of logistic regression models (Park, 2013). When the response variable has two categories (success or failure), binary logistic regression is used; when there are more than two response

variables, multinomial logistic regression is employed. A binary logistic regression for pipeline degradation, for example, comprises two response variables: 0 and 1. If the result is 0, the pipe is in bad condition; whereas, if the response variable is 1, the pipe is in good condition.

For a binary response variable Y and a single explanatory variable X, let $\pi(\mathrm{X}) = \mathrm{P}(\mathrm{Y} = 1 \mid \mathrm{X} = \mathrm{x}) = 1 - \mathrm{P}(\mathrm{Y} = 0 \mid \mathrm{X} = \mathrm{x})$, the logistic regression model has linear form for the logit of this probability as shown in Eq. 2.2 (Agresti, 2018).

$$\text{logit}\,[\pi(X)] = \log\left(\frac{\pi(X)}{1-\pi(X)}\right) = \alpha + \beta x \qquad \text{Eq.2.2}$$

Eq. 2.3 presents the formula for the probability $\pi(\mathrm{X})$, using the exponential function ($\exp(\alpha + \beta x) = e^{\alpha+\beta x}$).

$$\pi(X) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)} \qquad \text{Eq.2.3}$$

And the Eq. 2.4 presents the multiple logistic regression formula when multiple explanatory variables are used in the model (Agresti, 2018).

$$\log\left[\frac{\pi}{1-\pi}\right] = \log\left[\frac{P(Y=1 \mid X_1, X_2, \ldots, X_p)}{1-P(Y=1 \mid X_1, X_2, \ldots, X_p)}\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p \mathrm{X}_p = \alpha + \sum_{j=1}^{p} \beta_j \mathrm{X}_j \qquad \text{Eq.2.4}$$

where:

X_1, X_2, …, X_p are independent variables

α is the intercept parameter for category *i*

β is the regression coefficients

And the probability than Y=1 can be measured using an exponential transformation as shown in Eq. 2.5.

$$\mathrm{P}(\mathrm{Y} = 1 \mid X_1, X_2, \ldots, X_p) = \frac{e^{\alpha+\sum_{j=1}^{p} \beta_j \mathrm{X}_j}}{1+ e^{\alpha+\sum_{j=1}^{p} \beta_j \mathrm{X}_j}} \qquad \text{Eq.2.5}$$

An important parameter in logistic regression is odds ratio that measures the relationship between explanatory and response variables as shown in Eq. 2.6.

$$\frac{\pi(X)}{1-\pi(X)} = \exp(\alpha + \beta x) = e^{\alpha}(e^{\beta})^{x} \qquad \text{Eq.2.6}$$

Multinomial logistic regression is used when multiple levels of categorical response variables are in the model. Eq. 2.7 shows the multinomial logistic regression formula.

$$\log\left[\frac{\pi}{1-\pi}\right] = \log\left[\frac{P(Y=i \mid X_1, X_2, \ldots, X_p)}{1-P(Y=k \mid X_1, X_2, \ldots, X_p)}\right] = \alpha + \beta_{i1}X_1 + \beta_{i2}X_2 + \cdots + \beta_{ip}X_p =$$

$$\sum_{j=1}^{p} \beta_{ij}X_j \qquad \text{Eq.2.7}$$

where:

i = 1, 2, …, K-1 correspond to categories of the dependent variable

$X_1, X_2, \ldots, X_p$ are independent variables

α is the intercept parameter for category *i*

β is the regression coefficients associated with dependent category *i*

Using logistic regression, Ana et al. (2009) looked into the impact of sewage physical parameters on sewer pipeline structural degradation. The following characteristics were considered in this study: pipe age, size, depth, length, slope, form, material, sewer type, building time, and location. For selecting the predictor variables, they employed a backward stepwise regression approach. A Wald test and a likelihood-ratio test were used to determine the significance of the dependent variables. They also looked at how independent factors interact with one another. In a degradation model, the length of sewer pipes, for example, may be determined to be inconsequential, but when paired with another independent variable, it may become important. In this study, the age, material, and length of the sewers were determined to be important, and no validation procedure was utilized to validate the findings.

To forecast the structural state of sewage pipes, Kabir et al. (2018) used a Bayesian logistic regression model. The model was created by selecting 12,728 sewage mains from Calgary's wastewater network. The model was constructed using the following parameters: pipe age, material, diameter, length, slope, depth, rim elevation, up invert, and down invert. In this study, important factors were identified using the Bayesian model averaging approach, and sewage pipe condition was predicted using logistic regression. The significance of the independent variables was assessed using the P-test, Wald Test, likelihood-ratio test, and Durbin-Watson test. Sewer pipe conditions were classified into two groups: excellent and bad. A confusion test was used to verify the model's performance.

Malek Mohammadi (2019) used logistic regression to predict the condition of sewer pipes for Tamp city, Florida. According to the result of his model, the condition of 65.8% of sanitary sewer pipes was predicted correctly, however, pipes in conditions levels 2, 3, and 4 were not estimated properly by multinomial logistic regression (Malek Mohammadi, 2019).

2.5.3 Artificial Intelligence Models

2.5.3.1 Introduction

Artificial intelligence can be defined as “the study of mental faculties through the use of computational models” (Charniak and McDermott, 1985). In other definition, AI is “The art of creating machines that perform functions that require intelligence when performed by people” (Kurzweil, 1990).

According to Luger (2009), the artificial intelligence can be decomposed into several categories as describes in below items:

- Game playing
- Automated reasoning and theorem proving
- Expert systems
- Natural language understanding and semantics
- Modeling human performance
- Planning and robotics
- Languages and environments for AI
- Machine learning
- Alternative representations: neural nets and genetic algorithms
- AI and philosophy

In artificial intelligence models, the dependent variables are classified from a set of independent variables by learning from the available data. Artificial intelligence models can tackle complicated issues, and substantial research has been conducted in recent years to model infrastructure deterioration utilizing neural nets and machine learning approaches.

2.5.3.2 Neural Nets and Genetic Algorithms

The objective of developing neural nets and genetic algorithms is to provide a model which works parallel the structure of neurons in the human brain (Luger,2009). Among the neural net and genetic algorithm techniques, fuzzy set theory and neural networks (NNs) were used for modeling the deterioration of infrastructure facilities, (Flintsch and Chen 2004; Kleiner et al. 2004, Tran, 2007).

One of the methods used to forecast the degradation of sewage systems is Artificial Neural Networks. Najafi and Kulandaivel (2005) created a prediction model using an artificial neural network. The variables in this model were pipe length, size, material, age, depth, slope, and sewer type. The data was trained using backpropagation. The study revealed that using a neural network to construct a condition prediction model for pipelines is viable; however, model accuracy is strongly dependent on a bigger and more inclusive sample set.

Tran et al. (2007) created a neural network deterioration model to forecast the serviceability of underground stormwater pipelines. In this work, the model was calibrated using Markov Chain Monte Carlo simulation. In addition, the neural network's ranking performance was compared to several discrimination analysis models. This model included a number of independent factors such as pipe age, size, depth, slope, number of trees, pipe location, soil type, and wetness. According to the findings of the study, the performance of a neural network calibrated using the Markov chain approach outperforms that of a neural network calibrated with the backpropagation method.

Khan et al. (2010) created a structural condition prediction model to evaluate the significance and effect of key sewage pipe parameters. In this study, backpropagation and probabilistic neural networks were employed to convey the condition assessment of the pipes. The data for this model was contributed by the municipality of Pierrefonds, Quebec. The model was built using pipe material, diameter, depth, bedding material, length, and age. According to the created models, a neural network is capable of prioritizing inspection and restoration plans for existing sewage mains.

2.5.3.3 Machine Learning

Machine learning was characterized by Arthur Samuel in 1959 as a "Field of research that offers computers the ability to learn without being explicitly programmed" (Simon, 2013). By experimenting with

various prediction structures and algorithms, machine learning may learn directly from examples and experiences in the form of data (Bishop, 2016).

Machine learning can be classified in three broad categories based on the nature of learning as described below (Bishop, 2016):

- Supervised learning: in supervised learning models, the training data includes examples of input variables with their corresponding output variables.
- Unsupervised learning: application in which the training data comprises a set of input variables without any corresponding output variables.
- Reinforcement learning: same as unsupervised learning, the output variables are not given in the model and the targets should be predicted by trial and error.

Another classification of machine learning can be based on the desired output of the modeling systems. Below items define these categories:

- Classification: the outputs are divided into two or more classes and typically supervised learning are used to model this class.
- Regression: in this category the outputs are continuous rather than discrete and a supervised problem.
- Clustering: in clustering category, a set of inputs are classified into different groups. Unlike classification and regression, this is an unsupervised task.
- Density estimation: the distribution of inputs is found in some space in this category.
- Dimensionality reduction: simplifying the inputs by mapping them into a lower-dimensional space.

Machine learning is becoming increasingly popular in a variety of industries. In the wastewater business, machine learning models such as support vector machine (SVM), decision trees, random forest, and Bayesian regressions have been used to forecast sewage network damage.

Mashford et al. (2011) used a support vector machine to forecast sewage pipeline condition grade. The model's predictive performance was created using CCTV data obtained from the Adelaide wastewater collection network in South Australia. Sewer pipe condition was graded on a scale of 1 (excellent condition) to 5 (very bad condition). As input variables, pipe diameter, age, pipe location, slope, start invert, end invert,

material, soil type, soil corrosivity, grade, angle, sulfate soil, and groundwater level were employed. The study's findings revealed that the support vector machine has extremely strong prediction performance with 91 percent accuracy and may be utilized as a novel way to simulate sewage pipe degradation. The authors indicated that the study's drawback was a lack of adequate condition data.

To estimate the structural integrity of individual sanitary sewage pipes, Harvey and Mcbean (2014) employed a random forests model. The sewage database came from the city of Guelph in Ontario, Canada. The model was developed using several criteria such as pipe age, material, length, diameter, service type, slope, up elevation, down elevation, depth, land use, and pipe location. According to the findings, random forest models can accurately estimate the state of individual sewage pipes with an area under the ROC curve of 0.81. Random forest prediction models have the ability to cut project costs and time in half, and this method may be used to assess the status of uninspected sewage pipelines.

To forecast sewage pipeline condition ratings, Laakso et al. (2018) used random forest and binary logistic regression. Nonetheless, the elements that influence pipe degradation were explored in this study. The databases for this study were gathered in southern Finland. The model was created using a variety of predictors, including pipe age, diameter, material, slope, depth, length, soil type, road class, distance to tree, intersection with stormwater or water supply pipes, and yearly sewage flow. The models' accuracy was 62% for binary logistic regression and 67% for random forest, respectively. The study found that both logistic regression and random forest models may be utilized to forecast sewage pipeline condition in the future. In recent years, several sewer condition prediction models were developed. Table 2-3 shows detail of selected studies.

Table 2-3 Sewer Condition Prediction Models

	Authors	Year	Model	Number of Data
1	Davies et al.	2001	• Logistic regression	12,000
2	Ariaratnam et al.	2001	• Logistic regression	748
3	Micevski et al.	2002	• Markov chain	497
4	Najafi and Kulandaivel	2005	• Neural network	1050
5	Tran et al.	2006	• Neural network	583
6	Koo and Ariaratnam	2006	• Logistic regression	579
7	Tran et al.	2007	• Neural network • Multiple discrimination analysis	150
8	Chughtai and Zayed	2008	• Linear regression	-
9	Gat	2008	• Markov chain	5,262
10	Ana et al.	2009	• Logistic regression	1,316
11	Tran et al.	2009	• Neural network • Ordered probit model	417
12	Khan et al.	2010	• Neural network	200
13	Mashford et al.	2011	• Support vector machine	1,441
14	Salman and Salem	2012	• Ordinal regression • Logistic regression • Binary regression	11,373
15	Syachrani et al.	2013	• Decision tree • Neural Network	52,855
16	Sousa et al.	2014	• Neural network • Support vector machine • Logistic regression	745
17	Harvey and McBean	2014	• Random forest • Decision Tree • Support vector machine	1,825
18	Bakry et al.	2016	• Multiple regression	84
19	Gedam et al.	2016	• Linear regression	155
20	Kabir et al.	2018	• Bayesian logistic regression	12,728
21	Laakso et al.	2018	• Binary logistic regression • Random forest	6,700

22	Hernandez et al.	2018	• Logistic regression • Random forest	23,958 4,327
23	Malek Mohammadi	2019	• Logistic regression • K-nearest neighbors • XGBoost	20,282
24	Guzman et al.	2020	• Bayesian network	7,968
25	Mazumder et. al	2021	• KNN • Decision tree • Random forest • AdaBoost • XGBoost • LGBoost • CatBoost	959
26	Loganathan	2021	• Logistic regression • K-nearest neighbors • Random forest	32,751
27	Atambo	2021	• MLR • ANN	2,616

2.6 Chapter Summary

Pipe degradation is a complicated process, as detailed in earlier sections, and no one element may be the cause of pipe deterioration. Furthermore, wastewater agencies and municipalities are frequently short on funds to examine the status of all pipes in the network on a regular basis. As a result, an alternate method must be adopted to cut inspection costs while still providing a thorough plan for prioritizing and inspection scheduling. This chapter covered a variety of degradation models as well as elements that influence sewage pipe deterioration. However, condition prediction models for individual sewage pipes have yet to be thoroughly investigated, and the majority of research has concluded that novel data analysis methodologies can be used to estimate future sewer conditions and behavior.

Chapter 3 Prediction Model Concepts

3.1 Introduction

Prediction models for sewer pipes are influenced by the type and quantity of independent variables and the type of dependent variable. It is essential to choose a predictive model capable of predicting dependent variables with multiple categories since the dependent variable in this study is the condition rating for sewer pipes divided into five groups. Hence, the most suitable models for this dissertation were chosen based on the model's ability to predict multi-categorical dependent variables.

Logistic regression is the statistical model developed in this dissertation using SPSS software. The logistic regression model is the most popular for analyzing datasets containing two or more discrete outcome variables (Malek Mohammadi, 2019).

K-Nearest Neighbors (KNN) is the second model that was developed. The KNN model in this study is developed using Python, one of the most widely used programming languages in the data science field. Python was chosen for this study because it is open-source and has a wide range of free add-on libraries.

Tree-based models are developed as a third model. They are one of the most powerful learning techniques presented and are designed for classification problems. Python is used to develop these models too. Decision Tree, bagging approaches including Random Forest and boosting approaches including AdaBoost, Gradient Boosting Tree, and XGBoost are developed and discussed in this dissertation.

3.2 Logistic Regression

3.2.1 Introduction

Logistic regressions analyze the relationship between multiple independent variables and a categorical dependent variable. In logistic regression, the probability of occurrence of an event is estimated by fitting data to a logistic curve. The detail of binary and multinomial logistic regressions is explained in the following sections.

3.2.2 Binary Logistic Regression

Binary logistic regression is used to develop prediction models when the output (dependent or response) variable is binary. A binary are variables which only take two values. For example, the output of the model can be true or false, success or failure and zero or one (Malek Mohammadi, 2019). Eq. 3.1 presents the logistic regression formula when the dependent variable is binary (Hawari et al. 2017):

$$\log\left[\frac{\pi}{1-\pi}\right] = \mathrm{Y} = \left[\frac{\mathrm{p}(\mathrm{y}=1|\mathrm{x}_1 \ldots \mathrm{x}_\mathrm{p})}{1-\mathrm{p}(\mathrm{y}=1|\mathrm{x}_1 \ldots \mathrm{x}_\mathrm{p})}\right] = \alpha + \beta_1 \mathrm{x}_1 + \beta_2 \mathrm{x}_2 \ldots\ldots + \beta_\mathrm{p} \mathrm{x}_\mathrm{p} \quad \text{Eq. 3.1}$$

Where:

Y = dependent variable

$X_1, X_2, \ldots, X_p$ = independent variables

α = intercept parameter for category *i*

β = regression coefficients

And finally, the probability of y=1 can be measured using an exponential transformation as shown in Eq. 3.2.

$$\mathrm{P}(\mathrm{y} = 1 \,|X_1, X_2, \ldots, X_p) = \frac{e^{\alpha + \sum_{j=1}^{p} \beta_j \mathrm{x}_j}}{1 + e^{\alpha + \sum_{j=1}^{p} \beta_j \mathrm{x}_j}} \quad \text{Eq. 3.2}$$

π represents P (Y=1) meaning probability associated with outcome of condition 1. Consequently, 1-π represents P (Y=0) meaning probability of outcome of condition 0. π/(1-π) means the odds of having (Y=1).

Figure 3-1 illustrates logistic curve or logistic function which are used to estimate coefficient of the parameters in the model. In this example x varying from -4 to +4 while the y axes show the probability from 0 to 1.

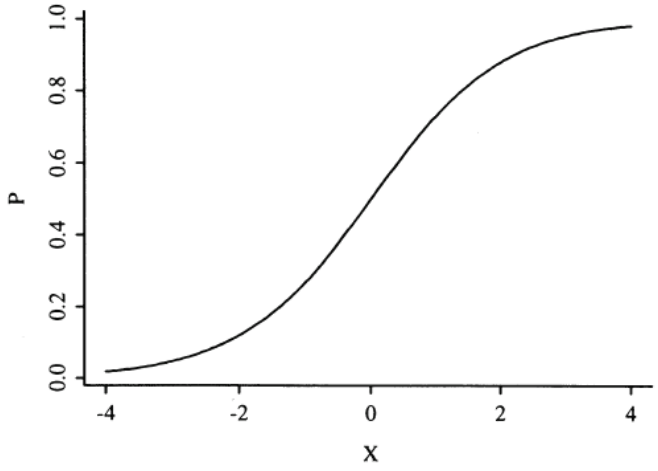

Figure 3-1 Logistic Function

(Harrell, 2016)

3.2.3 Multinomial Logistic Regression

When the dependent variable is categorical and has more than two levels, multinomial logistic regression can be used as an extension of binary logistic regression (Hawari et al. 2017). Assuming three conditions for a model, equations 3.3 and 3.4 represents multinomial logistic regression for a system with three condition levels 0, 1 and 2. Category zero (0) is used as the reference value. The objective of multinomial logistic regression in this case is to estimate the probability of having each of the three conditions and to convey the result in terms of odd ratio for choice of different conditions. Since, one of the categories is used as the reference value, two logit functions are required to develop the model. To develop the model, *p* covariate and a constant term denoted by the vector x (Hosmer et al., 2013).

$$\mathrm{g}_1(\mathrm{X}) = \log\left[\frac{P(Y=1 \mid X)}{P(Y=0 \mid X)}\right] = \beta_{10} + \beta_{11}X_1 + \beta_{12}X_2 + \cdots + \beta_{1p}\mathrm{X}_p \quad \text{Eq. 3.3}$$

$$\mathrm{g}_2(\mathrm{X}) = \log\left[\frac{P(Y=2 \mid X)}{P(Y=0 \mid X)}\right] = \beta_{20} + \beta_{21}X_1 + \beta_{22}X_2 + \cdots + \beta_{2p}\mathrm{X}_p \quad \text{Eq. 3.4}$$

Then, Probability of each condition levels can be calculated by Eq.3.5 through 3.7.

$$\mathrm{P}\,(\mathrm{Y}=0 \mid X) = \frac{1}{1+e^{g_1(X)}+e^{g_2(X)}} \quad \text{Eq. 3.5}$$

$$\mathrm{P}\,(\mathrm{Y}=1 \mid X) = \frac{e^{g_1(X)}}{1+e^{g_1(X)}+e^{g_2(X)}} \quad \text{Eq. 3.6}$$

$$\mathrm{P}\ \ (\mathrm{Y}=2 \mid X) = \frac{e^{g_2(X)}}{1+e^{g_1(X)}+e^{g_2(X)}} \quad \text{Eq. 3.7}$$

In the next sections more information will be given regarding significance of the models and variables.

3.2.4 Assumptions of Logistic Regression

The following are the assumptions of logistic regression (McDonald, 2009):

- Logistic regression does not assume that the independent variables are normally distributed.
- The observations should not come from repeated measurements.
- The correlation between independent variables should not be too high.
- The odds ratio and independent variables have a linear relationship.

3.2.5 Forward and Backward Stepwise Selection

The statistical techniques of forward and backward stepwise selection are used to screen the independent variables. In these methods, only variables with sufficient predictive power are retained in the

model, while unproductive variables are gradually removed. In forward stepwise, the intercept is chosen first, and then the variables that improve the model's performance are added sequentially. On the other hand, backward stepwise selection begins with the entire model and then deletes the variables that have the least influence. The variables with the lowest Z-scores should be removed from the model (Malek Mohammadi, 2019). A Z-score is a numerical measurement that describes a value's relationship to the mean of a group of values. Z-score is measured in terms of standard deviations from the mean. If a Z-score is 0, it indicates that the data point's score is identical to the mean score. A Z-score of +1.0 or -1.0 would indicate a value that is one standard deviation from the mean (Behboudian et al., 2008).

3.2.6 Odds Ratio

The odds ratio is a measure of association between an event occurring in one group, to the odds of it occurring in another group (Ana et al., 2009). For a probability of success π, the odds of success are given in Equation 3.8. The odds ratio is the ratio of the odds for x = 1 to the odds for x = 0 as given by the Equation 3.9 (Agresti, 2018).

$$odds = \frac{\pi}{1-\pi} \quad \text{Eq. 3.8}$$

$$OR = \frac{odds_1}{odds_0} = \frac{\left[\frac{\pi(1)}{1-\pi(1)}\right]}{\left[\frac{\pi(0)}{1-\pi(0)}\right]} = e^{\beta_1} \quad \text{Eq. 3.9}$$

The odds ratio is commonly used to estimated how much more likely or unlikely is the outcome to be present in groups where x = 1 or x = 0. Odds ratio greater than one shows that the outcome is most likely to occur when x = 1 and odds ratio less than one shows that the event is less likely to occur when x = 1 (Hosmer et al., 2013).

3.2.7 Significance of the Coefficients

According to Malek Mohammadi (2019), “Identifying the significant variables in the model is formulation and testing of a statistical hypothesis to determine if the independent variables are significantly related to the dependent variables. Typically, significance of the variables can be identified by comparing the observed dependent variables and predicted values after development of the model with and without independent variables. If the predicted values are more accurate by utilizing an independent variable in the model, then the variable is significant” (p. 90). Significance testing measures the strength of null hypothesis

by probability (the p-value). The significance is usually set to P-value ≤ 0.05 (95% confidence level). The null hypothesis assumes that coefficients (β) are zero. The alternative hypothesis assumes that not all the coefficients are equal to zero. Log-likelihood test and Wald test are the most common tests used in logistic regression to identify the significance of the variables

3.2.7.1 Log-likelihood Test

The log-likelihood function is used to compare the observed and predicted values. Equations 3.10 and 3.11 show the concept of log-likelihood function.

$$G = -2 \ln \left[\frac{(likelihood\ without\ the\ variable)}{(likelihood\ with\ the\ variable)} \right] \quad \text{Eq. 3.10}$$

$$G = 2 \left\{ \sum_{i=1}^{n} [y_i \ln \pi(X_i) + (1 - y_i) \ln(1 - \pi(X_i)] - [n_1 \ln(n_1) + n_0 \ln(n_0) - n \ln(n)] \right\} \quad \text{Eq. 3.11}$$

where $n_1 = \sum y_i$ and $n_0 = \sum(1 - y_i)$. For big samples, the G has a degree of freedom equal to the estimated number of parameters, and it follows a chi-square distribution (Harrell, 2016). A chi-square distribution is a continuous distribution with k degrees of freedom. It is used to describe the distribution of a sum of squared random variables. It is also used to test the goodness of fit of a distribution of data, whether data series are independent, and for estimating confidences surrounding variance and standard deviation for a random variable from a normal distribution (Kissel et al., 2017).

3.2.7.2 Wald Test

A method to determine the significance of the individual variables in logistic regression models is the Wald test. Equation 3.12 demonstrates that the Wald test is equal to the ratio between the maximum likelihood estimate and its standard error. This ratio has a normal distribution (Hosmer et al., 2013).

$$W_j = \left(\frac{\beta_j}{SE(\beta_j)} \right) \quad \text{Eq. 3.12}$$

where β_j is the coefficient of the predictor variable, and SE is the standard error of the coefficient. When an independent variable's Wald test result is zero, the variable is not significant and can be excluded from the model. In contrast, the variables should be part of the model if Wald is not zero.

3.2.8 Classification Table

The percentage of accurate predictions by the logistic regression models is displayed in classification tables. The outcomes of fitted logistic regression models are summarized in this table. A cut-

point (c) is established (the most common value is 0.5), and it is compared to each estimated probability in order to produce the discrete result of the classification table. They are placed in class one if the estimated probability exceeds the cut-point. In contrast, they are placed in the other groups if the estimated probability is below the cut-point (Hosmer et al., 2013).

3.3 k-Nearest Neighbors (k-NN)

3.3.1 Introduction

The k-NN algorithm is a machine learning method that can be applied to classification and regression issues. Both supervised and unsupervised learning approaches can use this model for prediction. The Nearest Neighbors method operates by locating the labels of K-nearest patterns in the data space.

To create a k-NN model, only the training dataset is needed. To classify a new data point, the algorithm locates the nearby data points in the training dataset, or its "nearest neighbors". Based on the majority of the K-nearest patterns in the data space, class labels are assigned for the unknown new data, x_j. A similarity measure based on the Minkowski metric is defined in equation 3.13 for data space (Kramer, 2016). The norm of the difference between two vectors is the distance between them in Minkowski metric.

$$\left\|\mathrm{x}' - \mathrm{x_j}\right\|^p = \left(\sum_{i=1}^{q} \left|(x_i)' - (x_i)_j\right|^p\right)^{1/p} \quad \text{Eq. 3.13}$$

In the case of binary classification with set of dependent variables $y = (1, -1)$, KNN is defined in Equation 3.14.

$$f_{KNN}(\mathrm{x}') = \begin{cases} 1 & if \ \sum_{i \in N_k(\mathrm{x}')} y_i \geq 0 \\ -1 & if \ \sum_{i \in N_k(\mathrm{x}')} y_i < 0 \end{cases} \quad \text{Eq. 3.14}$$

Where K is size of neighborhood with a set of $N_k(\mathrm{x}')$ of K-nearest patterns.

In its simplest configuration, the k-NN algorithm only examines one nearest neighbor, which is the training data point closest to the point we wish to classify (k=1). A voting technique labels the new data point of interest when more than one neighbor is taken into account (k>1). Voting is the total number of various class labels close to the data point of interest. The test data point will be placed in the class to which the majority of the class labels belong. Therefore, it is always advised to use an odd number for k to prevent

confusion when making predictions using nearest neighbors (Loganathan, 2021). Figure 3-2 illustrates the process for k=3.

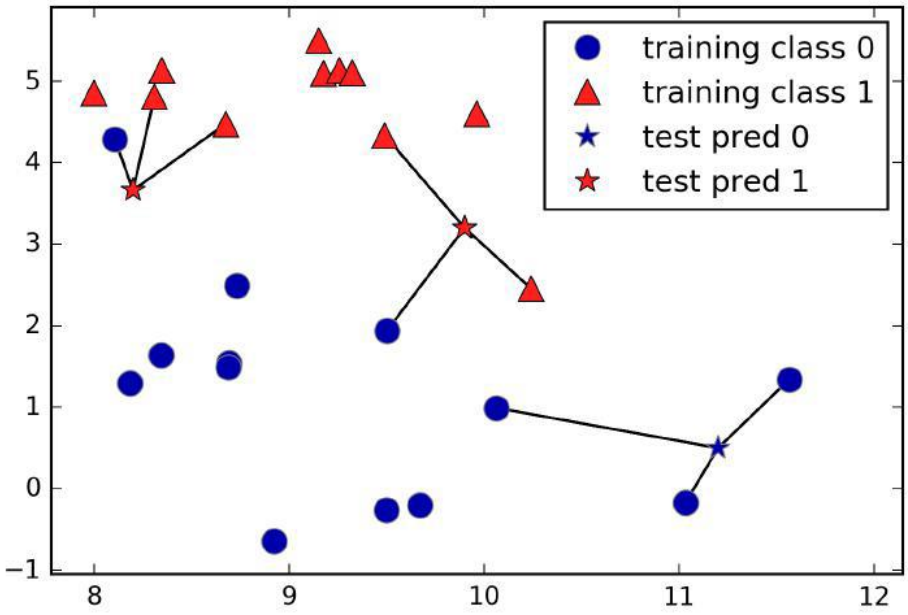

Figure 3-2 Predictions Made by the Three-Nearest-Neighbors Model

(Müller and Guido, 2016)

3.3.2 Evaluation of KNN

The confusion matrix, F-1 score, ROC curve, and area under the curve (AUC) were some of the methods used to assess the KNN model's performance. Section 3.5 provides an explanation of these metrics.

3.4 Tree-Based Models

3.4.1 Introduction

Tree-based models are popular for classification and regression problems. They are very powerful algorithms capable of fitting large datasets (Loh, 2014). They basically learn a hierarchy of if/else questions that leads to a decision. These questions are similar to those you might ask in a game of 20 Questions. Consider the following four animals mentioned in Figure 3-3. Your goal is to get to the right answer with as few if/else questions as possible (Müller and Guido, 2016).

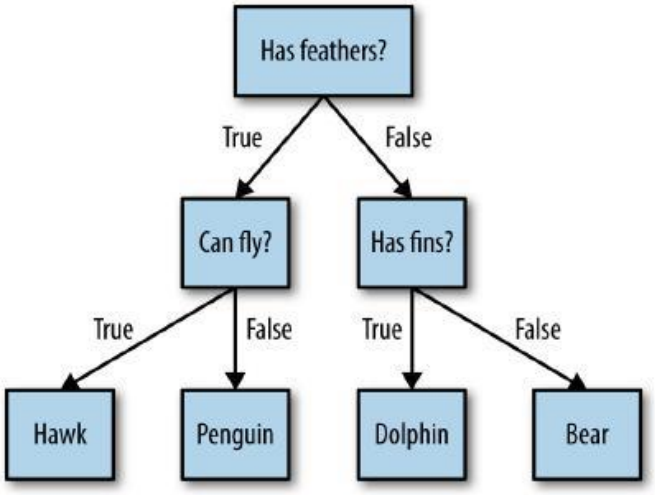

Figure 3-3 A Decision Tree to Distinguish Among Several Animals

3.4.2 Decision Tree

Before we get into how a decision tree works, we should know some key terms.

- Root node: The base of the decision tree.
- Splitting: The process of dividing a node into multiple sub-nodes.
- Decision node: When a sub-node is further split into additional sub-nodes.
- Leaf node: When a sub-node does not further split into additional sub-nodes; represents possible outcomes.
- Gini index: The Gini impurity index is one of the most widely used techniques for calculating the differences between the probability distributions of dependent variables. The Gini index calculates how often a random event is misidentified. As a result, a variable with a lower Gini index is preferable (Hastie et al., 2009). Gini index is calculated by equation 3.15 (Geron, 2017):

$$Gi\ (n) = 1 - \sum_{j=1}^{2} \left(p_j\right)^2 \qquad \text{Eq. 3.15}$$

The root node is the tree's base. A series of decision nodes flow from the root node, representing decisions to be made. Leaf nodes originate from the decision nodes to represent the consequences of those decisions. Each decision node represents a question or split point, and the leaf nodes that sprout from it represent possible answers. Leaf nodes sprout from decision nodes in the same way that a leaf sprouts from a tree branch. Figure 3-4 shows the elements of a decision tree.

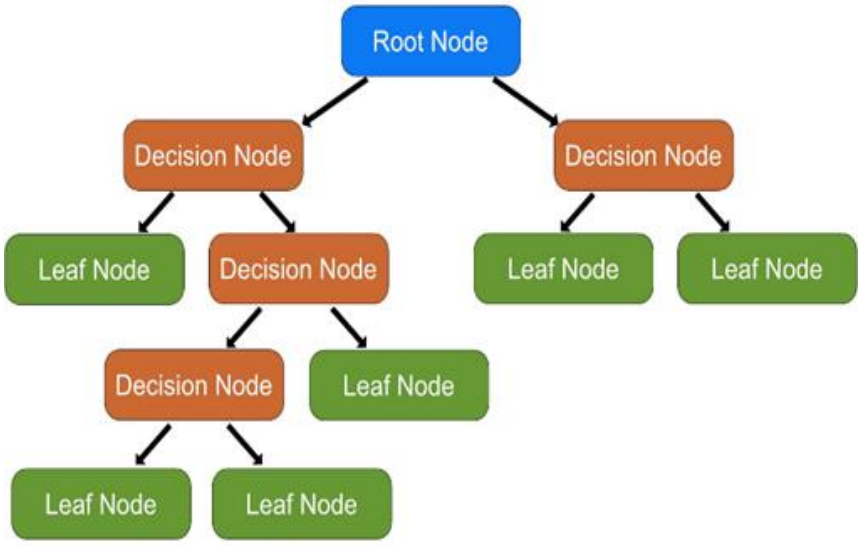

Figure 3-4 Elements of a Decision Tree

In this section, building a decision tree is explained by an example. Assume you find an iris flower and want to classify it. Let's look at how the tree in Figure 3-5 makes predictions (Geron, 2017).

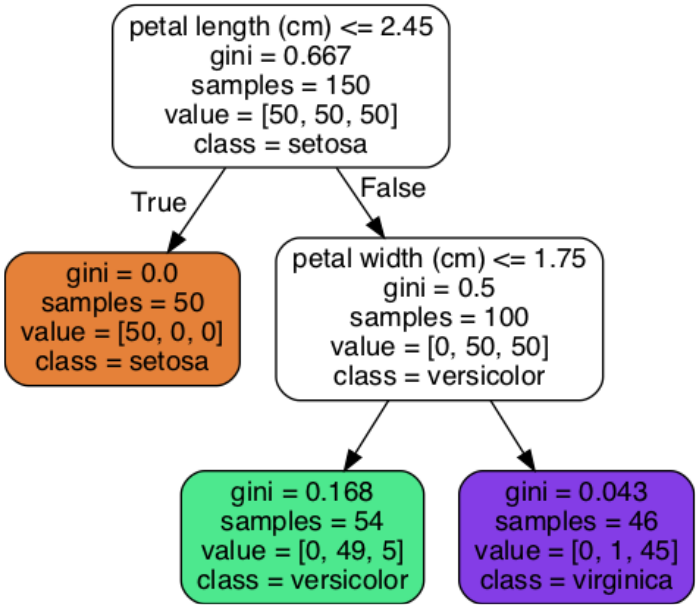

Figure 3-5 Iris Decision Tree

(Geron, 2017)

According to Geron (2017), “You start at the root node (depth 0, at the top): this node asks whether the flower’s petal length is smaller than 2.45 cm. If it is, then you move down to the root’s left child node

(depth 1, left). In this case, it is a leaf node (i.e., it does not have any children nodes), so it does not ask any questions: you can simply look at the predicted class for that node and the Decision Tree predicts that your flower is an Iris-Setosa (class=setosa). Now suppose you find another flower, but this time the petal length is greater than 2.45 cm. You must move down to the root's right child node (depth 1, right), which is not a leaf node, so it asks another question: is the petal width smaller than 1.75 cm? If it is, then your flower is most likely an Iris-Versicolor (depth 2, left). If not, it is likely an Iris-Virginica (depth 2, right). A node's samples attribute counts how many training instances it applies to. For example, 100 training instances have a petal length greater than 2.45 cm (depth 1,right), among which 54 have a petal width smaller than 1.75 cm (depth 2, left). A node's value attribute tells you how many training instances of each class this node applies to: for example, the bottom-right node applies to 0 Iris-Setosa, 1 Iris-Versicolor, and 45 Iris-Virginica. Finally, a node's gini attribute measures its impurity: a node is "pure" (gini=0) if all training instances it applies to belong to the same class. For example, since the depth-1 left node applies only to Iris-Setosa training instances, it is pure and its gini score is 0. Equation 3.15 shows how the training algorithm computes the gini score Gi of the i^{th} node. For example, the depth-2 left node has a gini score equal to $1 – (0/54)^2 – (49/54)^2 – (5/54)^2 ≈ 0.168$" (p. 179).

Overfitting the training data is one of the most common limitations of the decision tree. Overfitting is a modeling error in statistics that occurs when a function is too closely aligned to a limited set of data points. As a result, the model is useful in reference only to its initial data set and not to any other data sets (Müller and Guido, 2016). There is a common strategy to prevent overfitting in decision trees called pre-pruning. It includes limiting the tree's maximum depth and the number of leaves, but it is not always a solution for overfitting a decision tree. To overcome this problem, the random forest method is recommended.

3.4.3 Bagging

Utilizing a wide range of training algorithms is one technique to obtain a wide variety of classifiers. Another strategy is to train each predictor using the same training algorithm on various random subsets of the training data. Bagging (short for bootstrap aggregating) is the term used to describe sampling that includes replacement (Geron, 2017). Random Forest is a bagging approach.

3.4.3.1 Random Forest

A random forest is a collection of decision trees, where each differs slightly from the others. The theory behind random forests is that each tree may perform reasonably well at predicting, but will probably overfit on some portions of the data. If we construct numerous trees, each of which performs admirably and overfits differently, we can lessen the amount of overfitting by averaging the results (Müller and Guido, 2016). Specific guidelines are followed by RF for tree growth, tree combination, self-testing, and post-processing which are discussed in next chapter. Figure 3-6 shows the schematic of the random forest's concept.

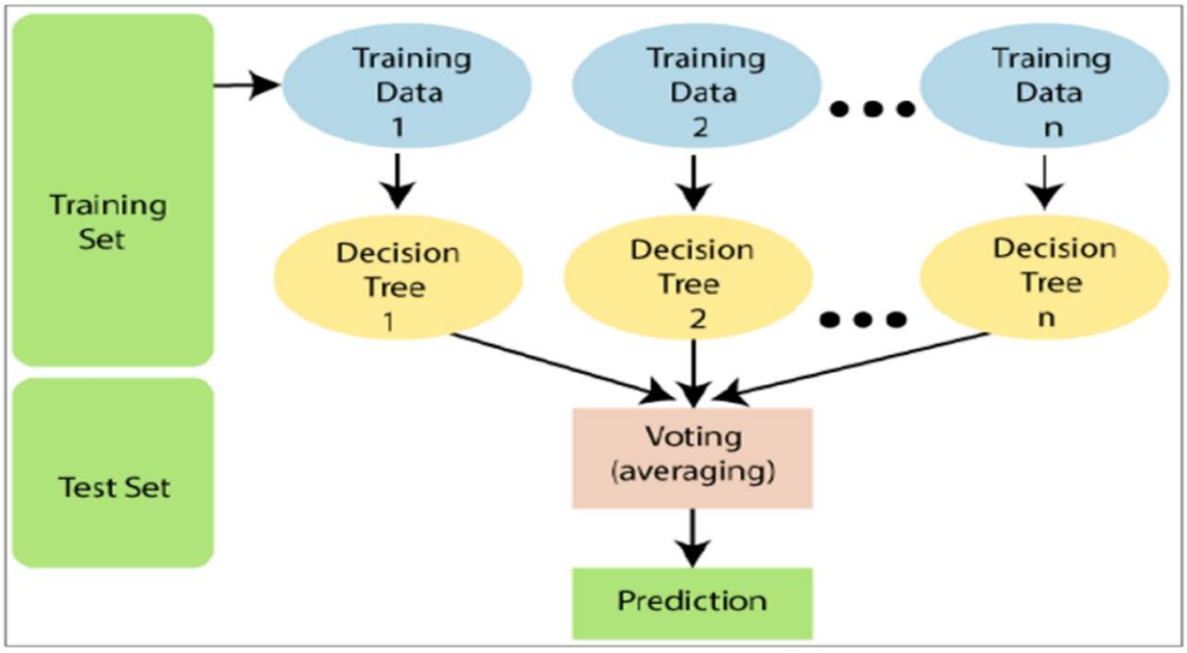

Figure 3-6 Structure of Random Forest Algorithm

(Loganathan, 2021)

The predictive power of classification variables is measured by the Gini index (Gi). GI is non-parametric, which means it is independent of the distribution of the data. RF is known to be more stable than other machine learning techniques in the presence of outliers and in high-dimensional parameter spaces, making it resistant to overfitting (Loganathan, 2021).

3.4.4 Boosting

The ensemble method combining several weak learners into a strong learner is called boosting. Predictors are trained successively using boosting approaches, with each predictor attempting to correct

the one before. Gradient Boosting, AdaBoost (short for Adaptive Boosting), and XGBoost (short for Extreme Gradient Boosting) are the most popular (Geron, 2017).

3.4.4.1 AdaBoost

Paying attention to the training instances that the predecessor underfitted is one technique for a new predictor to correct its predecessor. As a result, new predictors increasingly concentrate on the challenging scenarios. This is the method AdaBoost employs. For example, a first base classifier (such a Decision Tree) is trained and used to make predictions on the training set in order to develop an AdaBoost classifier. The relative weight of the incorrectly classified training instances is thus raised. The updated weights are used to train a second classifier, which then makes predictions on the training set while utilizing the updated weights, and so on (Geron, 2017). Figure 3-7 illustrates AdaBoost sequential training.

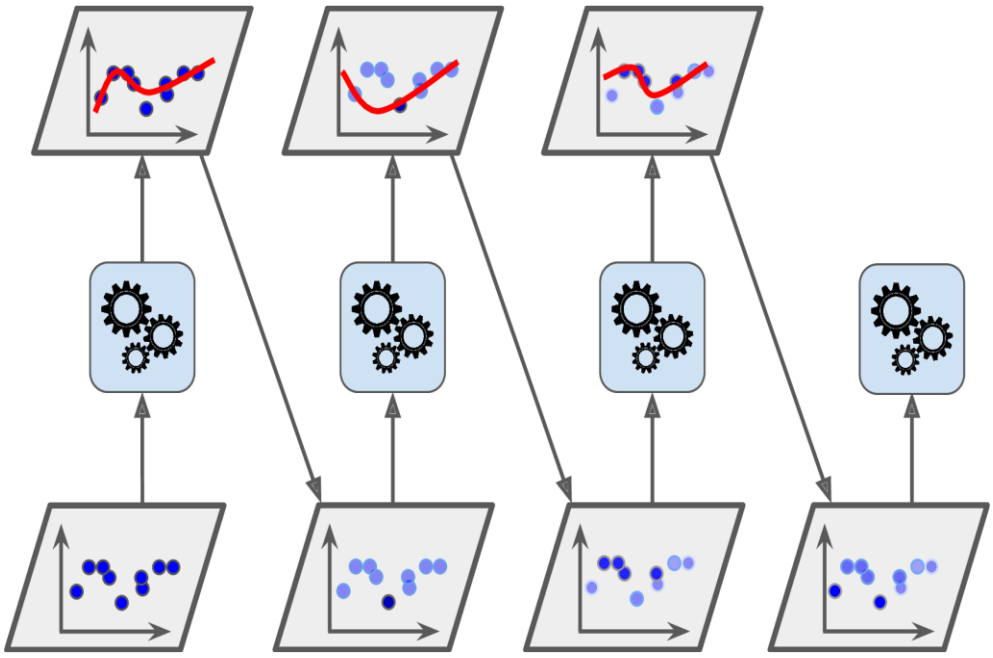

Figure 3-7 AdaBoost sequential training with instance weight updates

(Geron, 2017)

The learning rate of the boosting approach, in addition to pre-pruning and the number of trees in the ensemble, influences how strongly each tree tries to correct the errors of the preceding trees. Each tree can make greater corrections with a higher learning rate, enabling more complicated models (Müller and Guido, 2016). The decision boundaries for five consecutive predictors on the moons dataset are displayed

in Figure 3-8. Since the first classifier misclassifies many occurrences, their weights are increased. As a result, the second classifier performs better in these instances, and so on. The series of predictors is shown by the plot on the right, but the learning rate has been reduced by half (i.e., the misclassified instance weights are boosted half as much at every iteration). As you can see, AdaBoost gradually improves the ensemble by adding predictions to it.

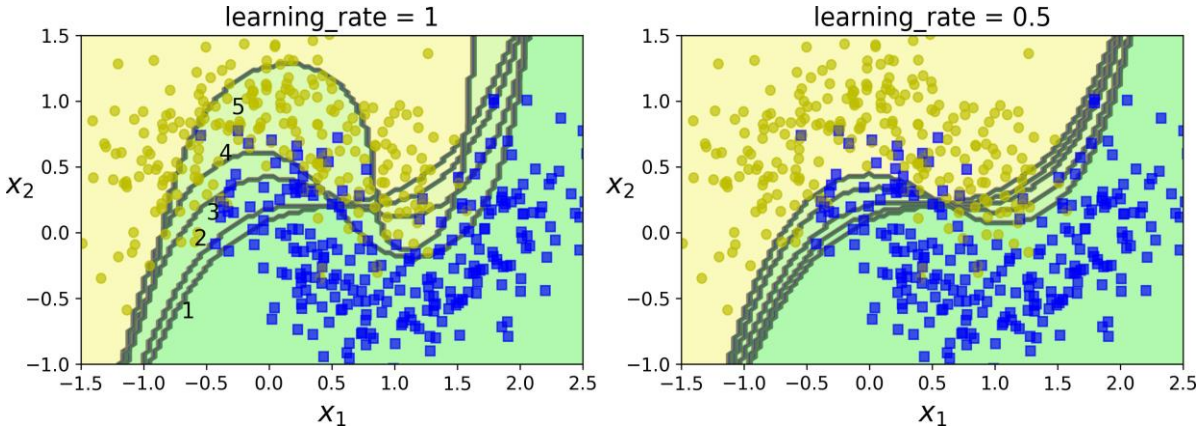

Figure 3-8 Decision boundaries of consecutive predictors

(Geron, 2017)

Equation 3.16 can be used to calculate the error rate of the training sample in the AdaBoost algorithm. In this equation, x is the independent variable and G(x) is a classifier that generates a prediction.

$$err = \frac{1}{N}\sum_{i=1}^{N} I\left(y_i \neq G(x_i)\right) \qquad \text{Eq. 3.16}$$

To create a succession of weak classifiers, weak classification algorithms are constructed. Next, a weighted classifier combination is used to create the final prediction, which is computed using Eq. 3.17 and displayed in Figure 3-9. The weights are successively assigned to each training observation, and the classification algorithm is then applied to the weighted observations. The weights for correctly predicted observations are dropped towards the end, while those for incorrectly categorized observations are increased (Malek Mohammadi, 2019).

$$G(x) = sign\left(\sum_{m=1}^{m} \alpha_m G_m(x)\right) \qquad \text{Eq. 3.17}$$

where the contributions of classifiers are weighted and $\alpha_1, \alpha_2, \ldots, \alpha_M$ are calculated using the boosting approach.

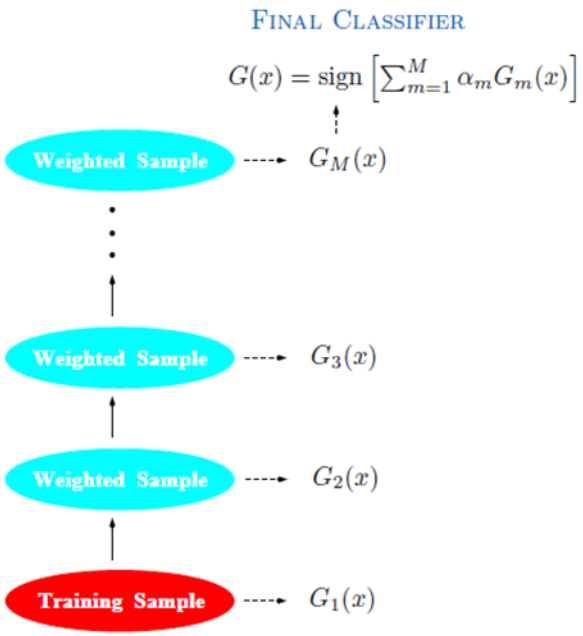

Figure 3-9 Schematic of AdaBoost Algorithm

(Hastie et al., 2009)

3.4.4.2 Gradient Boosting Trees

Gradient Boosting is a different widely used Boosting algorithm. Gradient Boosting operates similarly to AdaBoost by sequentially adding predictors to an ensemble, with each one correcting its predecessor. But unlike AdaBoost, which modifies the instance weights after each iteration, this approach aims to adapt the new predictor to the residual errors of the prior predictor (Geron, 2017).

Boosting models use a few fundamental operations to fit an additive expansion. The format of fundamental function expansions is presented in equation 3.18 (Malek Mohammadi, 2019).

$$(x) = \sum_{m=1}^{M} \beta_m \, b(x; \gamma_m) \qquad \text{Eq. 3.18}$$

Where $b(x; \gamma_m)$ are functions defined by a given set of parameters γ , and β_m are the expansion coefficients. In tree models, γ chooses the split variables and places the predictions at the terminal nodes and internal nodes, respectively.

A loss function is typically minimized to fit the gradient boosting trees, such as the squared-error or a likelihood-based loss function. The performance of the model's predictions is assessed using the loss function, a machine learning technique. The model is unsuitable for prediction when the loss function is high in value. The lower value of the loss function determines the model's potential to achieve greater

accuracy. It is, therefore, a technique to minimize the loss function to improve the performance of the models. The specifics of the boosting trees' loss function minimization are presented in Equation 3.19.

$$\min_{\{\beta_m,\gamma_m\}_1^M} \sum_{m=1}^{N} L(y_i, \sum_{m=1}^{M} \beta_m \, b(x;\gamma_m)) \quad \text{Eq. 3.19}$$

3.4.4.3 XGBoost

An optimized implementation of Gradient Boosting is available in the popular python library XGBoost, which stands for Extreme Gradient Boosting. It aims at being extremely fast, scalable, and portable. Suppose you want to apply gradient boosting to a large-scale problem. In that case, it might be worth looking into the XGBoost package and its Python interface, which is faster than the scikit-learn implementation of gradient boosting on many datasets. XGBoost computes second-order gradients, i.e., second partial derivatives of the loss function, which provide more information about the direction of gradients and how to get to the minimum of our loss function. This is the main difference between XGBoost with Gradient Boosting.

3.4.5 Feature Importance in Tree-Based Models

To summarize the operation of the tree, we can extract a few useful properties. Feature importance, which ranks the significance of each feature for the choice a tree makes, is the most popular one. Each variable is represented by a value between 0 and 1, where 1 indicates that the variable completely predicts the target and 0 indicates that it is not used at all. Features' weight values always add up to 1 (Geron, 2017).

According to Biau and Scornet (2016), to assess the significance of the variables, two measures of significance are typically employed: Mean Decrease Impurity (MDI) and Mean Decrease Accuracy (MDA). MDI counts the number of times an independent variable is used to divide a node. The weighted decrease of impurity from splitting on the variable, averaged over all trees, is measured using the mean decrease impurity (MDI). On the other hand, Mean decrease accuracy (MDA) is based on averaging the variation in out-of-bag error estimation between before and after the permutation over all trees.

3.4.6 Evaluation of Tree-Based Models

The goal of supervised learning techniques, which are often trained by a set of data, is to create a model that can make predictions. To determine the model's quality and to identify its key parameters, the performance of prediction models must always be assessed. There are numerous methods for assessing

the effectiveness of machine learning models. In this dissertation, tree-based models are assessed using a confusion matrix, ROC curve, and area under the curve (AUC). These metrics are discussed in section 3.5.

3.5. Evaluation Metrics

Different evaluation measures are covered in detail in this section of the dissertation. The type of expected output from the classification model would be used to guide the selection of a specific measure.

3.5.1 Confusion Matrix

The trained classifier's assigned class is compared to each test sample's actual class in a confusion matrix. The confusion matrix calculates the proportion of elements for each class that was correctly or wrongly predicted. The examples in the model that are successfully classified are determined by true positive or true negative (TP/TN). As shown in Figure 3-10, false positive or false negative (FP/FN) instances, on the other hand, indicate positive or negative examples that are wrongly categorized (Hossin and Sulaiman, 2015; Malek Mohammadi, 2019).

Predicted

Actual Class		Predicted positive	Predicted negative
	Positive Instances	True positive (TP)	False negative (FN)
	Negative Instances	False positive (FP)	True negative (TN)

Figure 3-10 Confusion Matrix

In a confusion matrix, the correctly classified items are arranged on the major diagonal from top left to bottom right, and their placement corresponds to the proportion of instances the two classes agree. In the confusion matrix above, TP (True Positive) stands for cases that were truly predicted to be positive, and TN (True Negative) stands for instances that were truly anticipated to be negative. False negative elements, or FNs, are those that the model predicted as negative but are actually positive, and false positives, or FPs, are those that the model forecasted as positive but are actually negative. For a model to perform better, it

could be noted that the amount of elements in cells other than the primary diagonal cells should be kept to a minimum (Loganathan, 2021). Based on the values in Figure 3-9, the measurements shown in Table 3-1 can be calculated in the confusion matrix method (Hossin and Sulaiman, 2015):

Table 3-1 Metrics Extracted from Confusion Matrix

(Hossin and Sulaiman, 2015)

Metrics	**Formula**	**Evaluation Focus**
Accuracy	$\frac{TN + TP}{TP + FN + FP + TN}$	In general, the accuracy metric measures the ratio of correct predictions over the total number of instances evaluated.
Misclassification Rate (Error Rate)	$\frac{FN + FP}{TP + FN + FP + TN}$	Misclassification error measures the ratio of incorrect predictions over the total number of instances evaluated.
True Positive Rate (Sensitivity) (Recall) (R)	$\frac{TP}{TP + FN}$	This metric is used to measure the fraction of positive instances that are correctly classified.
False Positive Rate	$\frac{FP}{FP + TN}$	This metric is used to measure the fraction of positive instances that are wrongly classified.
Precision (P)	$\frac{TP}{TP + FP}$	Precision is used to measure the positive instances that are correctly predicted from the total predicted instances in a positive class.
True Negative Rate (Specificity)	$\frac{TN}{TN + FP}$	This metric is used to measure the fraction of negative instances that are correctly classified.
False Negative Rate	$\frac{FN}{FN + TP}$	This metric is used to measure the fraction of negative instances that are wrongly classified.
F-1 Score	$\frac{2\,(R)(P)}{(R + P)}$	This metric represents the harmonic mean between recall and precision values.

3.5.2 ROC Curve and AUC

The Receiver Operating Characteristic (ROC) curve depicts the ratio of true positive to false positive rates. The true positive rate is shown on the Y-axis of a ROC curve, while the false positive rate is

shown on the X-axis. The best point on the ROC curve is reached when the model correctly predicts all positive examples. As a result, the model has higher overall accuracy when the ROC curve is closer to the upper left corner. An example ROC curve can be shown in Figure 3-11 (Malek Mohammadi, 2019).

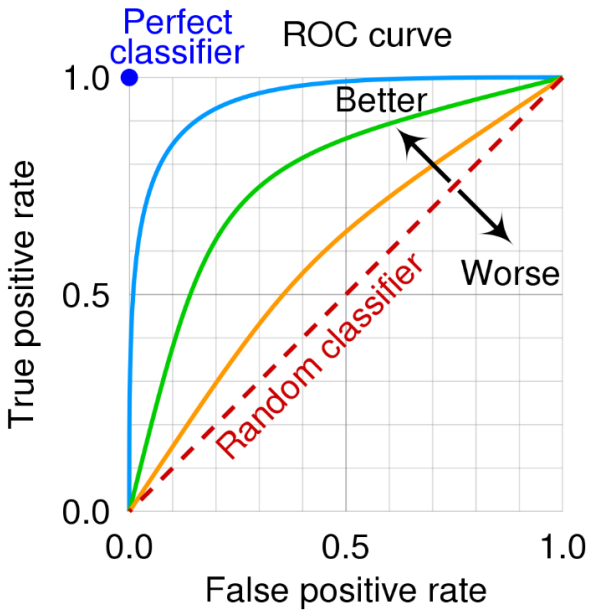

Figure 3-11 ROC Curve

The colored area under the ROC curve in Figure 3-12 is the area under the curve (AUC), another interesting measurement taken from the ROC curve. The AUC spans a range of 0 to 1, as the chart's dimension is a unit square. If the AUC were larger, it may be concluded that the model would perform better in predictions. The probability that a classifier would rank a randomly selected positive instance higher than a randomly selected negative instance is known as a classifier's AUC (Loganathan, 2021).

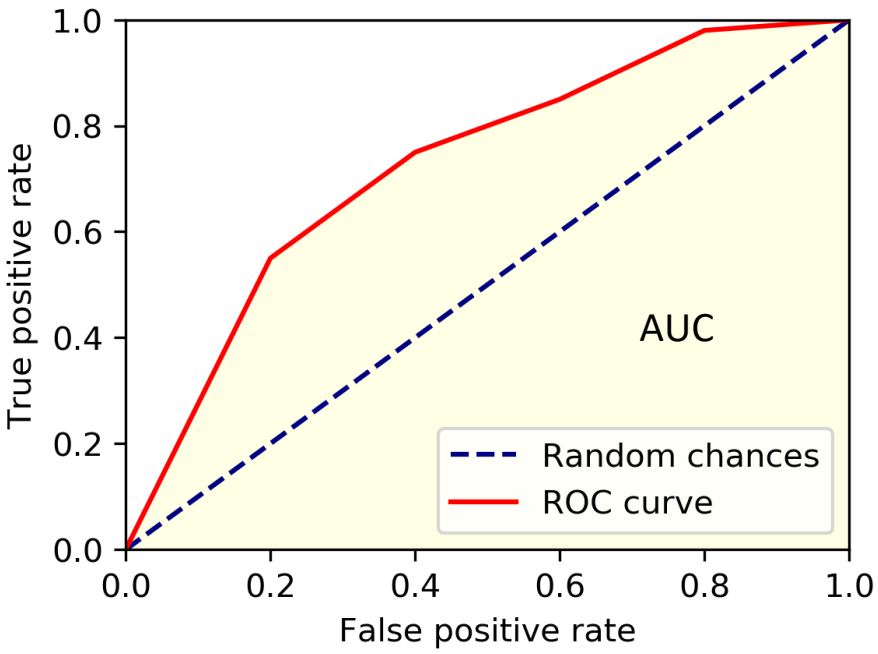

Figure 3-12 Area Under Curve (AUC)

3.5.3 Precision

According to Equation 3.20, the precision is the percentage of actually positive anticipated positives. According to the definition, a model's precision is important when the need for accuracy in the forecast is important. In other words, when one class of the output variable has rarer occurrences than the other class, the precision of a model is critical. Precision (P) would be a key evaluation parameter during model selection as the PACP score of 5 has proportionally fewer instances than other classes, making it more significant for correct prediction (Hossin and Sulaiman, 2015; Loganathan, 2021).

$$p = \frac{TP}{TP+FP} \qquad \text{Eq. 3.20}$$

3.5.4 Recall

The ability of a model to capture every positive element in the dataset is measured by its Recall (R). According to Equation 3.21, Recall can be defined as the proportion of true positive elements to all positively classified elements. It is clear that Recall serves as a measure of the model's predictive power for the positive class. For instance, after training, a model should be able to identify all pipe segments having a PACP score of 5 (Hossin and Sulaiman, 2015; Loganathan, 2021).

$$R = \frac{TP}{TP+FN} \qquad \text{Eq. 3.21}$$

3.5.5 F1-Score

F1-score is generated by calculating the harmonic mean of precision and recall and combining them into a single measure to evaluate the effectiveness of the classification model, as indicated in Equation 3.22. The F1-score is measured on a scale from 0 to 1, with 1 indicating greater model performance and 0 indicating worse performance. Both precision and recall contribute equally to the F1-score because it is a weighted average of the two. As a result, it can be used to determine the best way to trade off the two values. F1-score is discovered to be an essential metric to assess the effectiveness of a developed model based on the evaluation criteria (Hossin and Sulaiman, 2015; Loganathan, 2021).

$$F1 = \frac{2\,(R)(P)}{(R+P)} \qquad \text{Eq. 3.22}$$

3.5.6 Confusion Matrix for a Multi-Class Classification

All of the measures mentioned so far have been based on a binary classification confusion matrix, which should be noted. The pipe condition must be predicted in this study among five different classifications; binary classification cannot be used. Two alternative F1-scores were generated to evaluate multi-class classification models that must consider all classes: Micro F1-score and Macro F1-score (Grandini et al., 2020; Loganathan, 2021). The multi-class confusion matrix was used to assess multiple precision and recall for various classes to include all of the classes in the F1 score. Figure 3-13 illustrates a multi-class classification confusion matrix as an example.

Predicted Class

Actual Class	1	2	3	4	5
1	TN	TN	TN	FP	TN
2	TN	TN	TN	FP	TN
3	TN	TN	TN	FP	TN
4	FN	FN	FN	TP	FN
5	TN	TN	TN	FP	TN

Figure 3-13 A Five-Class Confusion Matrix

The confusion matrix depicted in the preceding figure has five output classes: 1, 2, 3, 4, and 5. Metrics are calculated using the confusion matrix, one class of interest at a time. Class 4 is regarded as the target class of interest in Figure 3-12, for example. Therefore, TP is equal to the number of class 4 components successfully anticipated. FP and FN represent elements that are mistakenly categorized along row and column of class 4, respectively, just like a binary confusion matrix. Furthermore, last, TN is used to describe all other remained cells. Quantities are recalculated when an interest class is altered, and the confusion matrix cell labels are adjusted accordingly (Grandini et al. 2020; Loganathan, 2021).

There are two types of F1-Scores: Micro and Macro. Equation 3.23's Micro-F1 score formula reveals that it is similar to the accuracy formula. The majority of classes will be given more weight because the approach considers the dataset as a whole. Therefore, it may be inferred that the micro F1-score is not a suitable metric for this study.

$$\frac{\sum_{k=1}^{K} TP_k}{\text{Grand Total}} \qquad \text{Eq. 3.23}$$

By computing the macro-average precision and recall for each target class, the macro F1-Score is estimated. Equations 3.24 and 3.25 illustrate how macro-average precision and recall are directly computed as the arithmetic mean of the same for individual classes. Equation 3.26 demonstrates that the macro F1-score is the harmonic mean of macro average precision and macro average recall (Grandini et al., 2020; Loganathan, 2021).

$$\text{Macro-Average Precision} = \frac{\sum_{k=1}^{K} \frac{TP_k}{TP_k + FP_k}}{K} \qquad \text{Eq. 3.24}$$

$$\text{Macro-Average Recall} = \frac{\sum_{k=1}^{K} \frac{TP_k}{TP_k + FN_k}}{K} \qquad \text{Eq. 3.25}$$

$$\text{Macro F1-Score} = 2 * \left(\frac{\text{Macro Precision} * \text{Macro Recall}}{\text{Macro Precision} + \text{Macro Recall}} \right) \qquad \text{Eq. 3.26}$$

The numerators are numbers between 0 and 1 according to the macro-average precision and recall formulas. This demonstrates that classes of various sizes are equally weighted and that the measure is unaffected by class size. In other words, the minority class will be given the same weight as the majority class. Therefore, it may be concluded that high Macro F1-score values show that the trained model performs well across all classes, but low Macro F1-score values show that classes are poorly predicted by the trained model (Loganathan, 2021). Consequently, Macro F1-score could be regarded for this study as an essential parameter for assessing the model's performance. Figure 3-14 explains calculating the Macro-Average F1-Score of a 3-classes confusion matrix.

Predicted Classification

Actual Classification	Classes	1	2	3	*Recall (r)*
	1	a	b	c	$\frac{a}{a+b+c}$
	2	d	e	f	$\frac{e}{d+e+f}$
	3	g	h	i	$\frac{i}{g+h+i}$
	Precision (p)	$\frac{a}{a+d+g}$	$\frac{e}{b+e+h}$	$\frac{i}{c+f+i}$	

Figure 3-14 Calculating the Macro-Average F1-Score for a 3-Classes Confusion Matrix

After calculating precision and recall for each class, as shown in the figure above, Macro-Average Precision (P), Macro-Average Recall (R), and Macro-Average F1 are calculated as below:

$$P = \frac{\frac{a}{a+d+g} + \frac{e}{b+e+h} + \frac{i}{c+f+i}}{3}$$

$$R = \frac{\frac{a}{a+b+c} + \frac{e}{d+e+f} + \frac{i}{g+h+i}}{3}$$

Macro-Average F1 = $\frac{2RP}{R+P}$

3.6 Chapter Summary

This chapter thoroughly studied the details of KNN models, tree-based models, and logistic regression. The discussions in this chapter confirmed that statistical and artificial intelligence models could be used as classifiers to determine the state of sanitary sewer pipes. Also, various evaluation metrics for different modeling approaches were addressed. The next chapter will outline the origin of the sanitary sewer pipe database and the various data preparation steps.

Chapter 4 Data Collection, Preparation, and Analysis

4.1 Introduction

This chapter goes over data collection, preparation, and processing. Histograms displaying the frequency of variables are provided. They are used to compare the factors influencing sewer pipe conditions. Data descriptive statistics and correlation analysis are presented too.

This study is based on the combined data collected from the Dallas Water Utilities Wastewater Collection System (Dallas, TX) and the City of Tampa's Wastewater Department (Tampa, FL). The purpose of combining two separate datasets was to have more diverse data about sewer pipes and their environmental conditions to develop a more comprehensive model. Also, increasing the data to reach a more accurate model was another reason for combining the datasets.

CCTVs are widely used in the United States to inspect sewer pipes (NASSCO, 2018). They are employed in both cities' sanitary sewer pipes' inspection and condition assessment process. Pipeline Assessment and Certification Program (PACP) guidelines were used to evaluate the condition of pipes in both cities on a scale of 1 to 5, with 1 indicating a pipe with no or few defects and 5 indicating failing conditions. The inventory of the sewer system is stored using geographic information system (GIS) databases. The recorded database inventory includes information such as pipe installation details, pipe location in reference to geographical maps, and so on.

4.2 Dataset Preparation

The data acquired from Dallas Water Utilities (DWU) contained a total number of 3,376 individual manhole to manhole pipe segments. This dataset included a given unique name for future identification of pipe segment referred to as DWU_Key, size, installation and inspection date, material, slope, depth, length, surface condition, soil pH, soil type surrounding the pipe, and PACP score (Atambo, 2021; Atambo et al., 2022). Figure 4-1 shows the location of sewer pipes in Dallas.

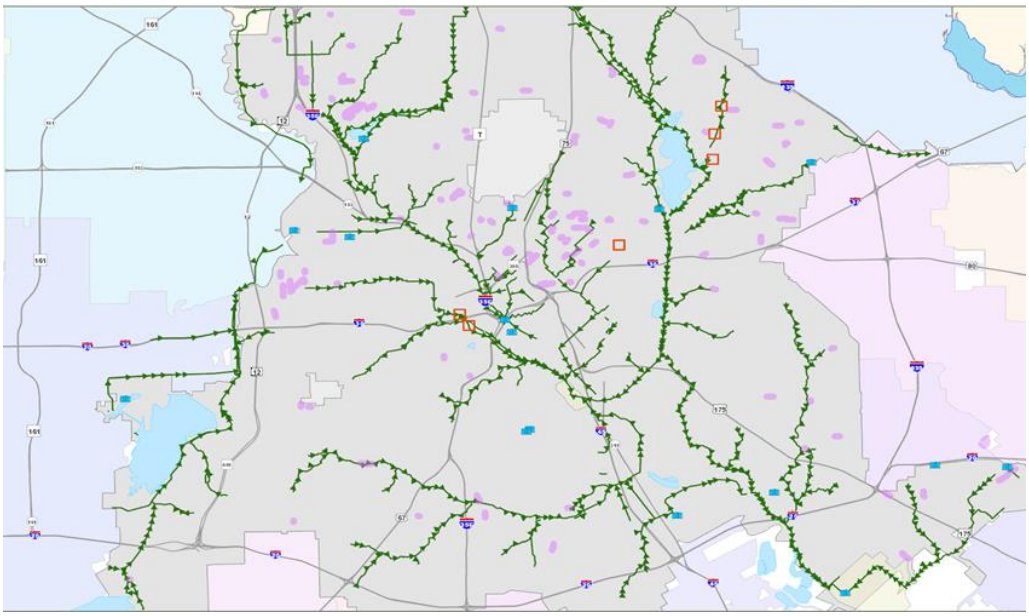

Figure 4-1 Sewer Pipe Network of City of Dallas

(Dallas Water Utilities)

One of the important steps in preparing data for further analysis is pre-processing. The collected data is processed to avoid any null values as part of the data preparation process. The dataset's null values were excluded from further analysis. Also, some pipes with minority materials such as asbestos-cement (AC), cast iron (CI), ductile iron pipe (DI), prestressed concrete cylinder pipe (PCCP), and High-density polyethylene pipe (HDPE), etc. were excluded to avoid any misclassification or error during model. Individual features such as age were calculated based on collected data as the next step in data preparation to include in the model development phase. Ultimately, the surface condition of each pipe segment (the road type that the pipe is buried beneath it) was identified using GIS and added to the dataset. Figures 4-2 and 4-3 show the original dataset of DWU and pipe locations, respectively.

DWUKEY	DIAMETER	INSTALLDATE	LOCATIONDESCRIPTION	MATERIAL	SLOPE	Depth	Shape_Length	PH_AVG	SOIL_TEX	PACP_SCORE
139587	8	01-Jan-34	01094080002M-TMP1629928M	UNK	0.4	5	602.3725565	6.5	Sand	3 - Fair
876728	8	11-Apr-00	09019020028N-09019020027M	PVC	1.85	8	136.897473	7.5	Loam	1 - Excellent
1454949	8	13-Aug-07	TMP1454943M-TMP1586177M	PVC	0.33	4	347.4813611	7.5	Clay	1 - Excellent
1672700	12	08-Nov-13	TMP1672695M-TMP1672694M	PVC	0.2	5	632.749228	7.9	Clay	1 - Excellent
1265040	8	02-Apr-04	01014030065M-01014030145M	PVC	4.2	5	399.7080588	8.2	Loam	1 - Excellent
157675	6	15-Jul-81	16049080002M-16049080001M	PVC	1.6	10	390.2358047	7.9	Clay	1 - Excellent
115698	6	24-Feb-75	05092010003C-05092010001M	VCT	0.6	5	613.4049029	7.9	Clay	1 - Excellent
984259	8	17-Sep-91	35027050051M-35027050050M	PVC	1.73	4	272.1759355	8.2	Clay	1 - Excellent
1312187	8	20-Nov-04	25066000007M-25066000004M	PVC	1.8	10	494.9968642	8.2	Clay	1 - Excellent
1258820	8	22-Jan-04	04018220016N-04018220015M	PVC	1	7	272.2896879	8.2	Clay	1 - Excellent
1485836	8	17-Mar-09	TMP1485829M-TMP1485827M	MULT	0.61	5	458.7938036	7.9	Clay	1 - Excellent
1154492	10	28-Jun-93	12066020005M-12066020004M	PVC	0.3	9	87.19764267	7.9	Clay	1 - Excellent
890355	8	01-Oct-58	04022000210M-04022000200M	CONC	0.8	6	238.2963048	7.9	Clay	2 - Good
975705	6	12-Oct-59	17007180001M-17007000170M	VCT	0.6	8	318.4237443	7.9	Clay	2 - Good

Figure 4-2 Original Dataset of Dallas Sewer Pipes

Figure 4-3 Location of Sewer Pipes in Dallas City (Dallas Water Utilities)

As a result, the final dataset for Dallas City containing 3,104 data points was used for analysis and model development. In the Dallas City dataset, the soil types were Sand, Loam, Clay, and Rock, and the pipe materials were Polyvinyl Chloride (PVC), Vitrified Clay Pipe (VCP), and Reinforced Concrete (RC).

In case of Tampa City dataset, the sewer inventory dataset included 5,144 manholes to manhole pipe segments. To make it easier to track individual pipes, each pipe segment was assigned a unique number (Facility ID). In addition, the dataset came with a shape file (GIS file). The dataset included pipe attributes such as installation date, material, diameter, length, depth, down elevation, up elevation, and location. Figure 4-4 illustrates location of sewer pipes in Tampa city.

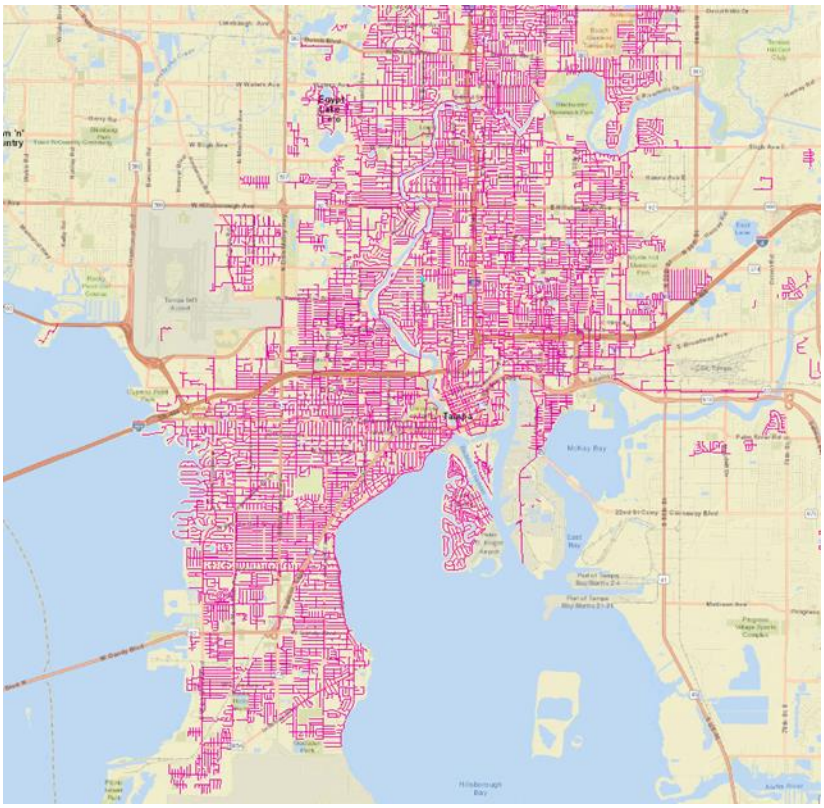

Figure 4-4 Sewer Pipe Network of Tampa City

As a first step, pipes with missing information on pipe installation date, depth, material, length, and condition scales were excluded from the dataset. They were around 2,000 segments. Then, pipe's ages and slopes were calculated based on available data. In the next step, pipe materials with a low population in the dataset, such as ductile iron, reinforced concrete, and plastic pipes, were removed. The total number of all these pipes was approximately 200. Also, using the mentioned shape file, soil data of pipes, including

soil pH and type of the soil surrounding pipe segments and surface conditions, were extracted. Figure 4-5 illustrates the combination of sewer pipe's location and soil datasets in GIS.

Figure 4-5 Combination of Sewer Network and Soil Dataset

Finally, the condition rating of each pipe was obtained from the dataset. The overall condition of pipes existed in this database in addition to some information such as pipe rating, quick rating, and pipe rating index for structural and operational conditions. The final dataset of Tamp City contains 2,944 individual pipe segments. The soil type in Tampa city included: Sand and Gravel, and the pipe material was Polyvinyl Chloride (PVC) and Vitrified Clay Pipe (VCP).

Finally, two datasets were combined into one set, and boxplot technic were used to remove outliers from the dataset. Outliers numerically distant from the rest of the data are frequently found in observed datasets. Outliers are typically larger or smaller than the observed values in the dataset. A boxplot is a graphical tool for displaying the variation of continuous data. The boxplot identifies the median, lower quartile, upper extreme, and upper extreme. Table 4-1 shows that the final dataset contains 4,803 individual pipe segments with various physical and environmental parameters.

Table 4-1 Variables Included in Sewer Pipe Dataset

Category	Variable	Description
Physical	Age (years)	Time difference between the installation date of the pipe segment and the date of inspection
	Material	Type of pipes material (PVC, VCP, and RC)
	Diameter (in)	Diameter of the pipe segment
	Depth (ft)	Depth of overburden above the pipe segment
	Slope (%)	Vertical displacement of the pipe segment per horizontal displacement
	Length (ft)	Length of the pipe segment between two manholes
Environmental	Soil Type	Type of soil surrounding the pipe (Sand, Gravel, Loam, Clay, and Rock)
	Soil pH	A numerical expression of the relative acidity or alkalinity of a soil sample
	Pipe Location	Category of the ground surface where the pipe is located (Highway, Street, Alley, and Easement)

The final dataset for analysis consisted of nine independent variables and one dependent variable. Table 4-2 shows the details of each variable. They are classified based on their statistical type.

Table 4-2 Statistical Types of Variables in Dataset

Variable Type	Variable	Data Type
Independent	Age	Continuous Numerical
	Length	
	Slope	
	Diameter	
	Depth	
	Soil pH	
	Material	Nominal Categorical
	Soil Type	
	Pipe Location	
Dependent	Condition Rating	Ordinal Categorical

4.3 Exploratory Data Analysis

4.3.1 Age

The age of a sewer pipe segment is determined by the difference between the inspection and installation dates. Figure 4-6 shows the distribution of sewer pipe age. The dataset contains pipes ranging in age from 1 to 121 years. Based on Figure 3-6, nearly 2% of the pipes are less than ten years old, 2% are over 80 years old, and the rest is between 10 and 80 years old. The maximum quantity is for 60-70 years old pipes (23.18%). Figure 4-7 depicts the age state concerning condition rating. According to Figure 4-7, the average age of pipes with condition rating 1 is around 38 years, the average age of pipes with condition rating 5 is around 63 years, and so on.

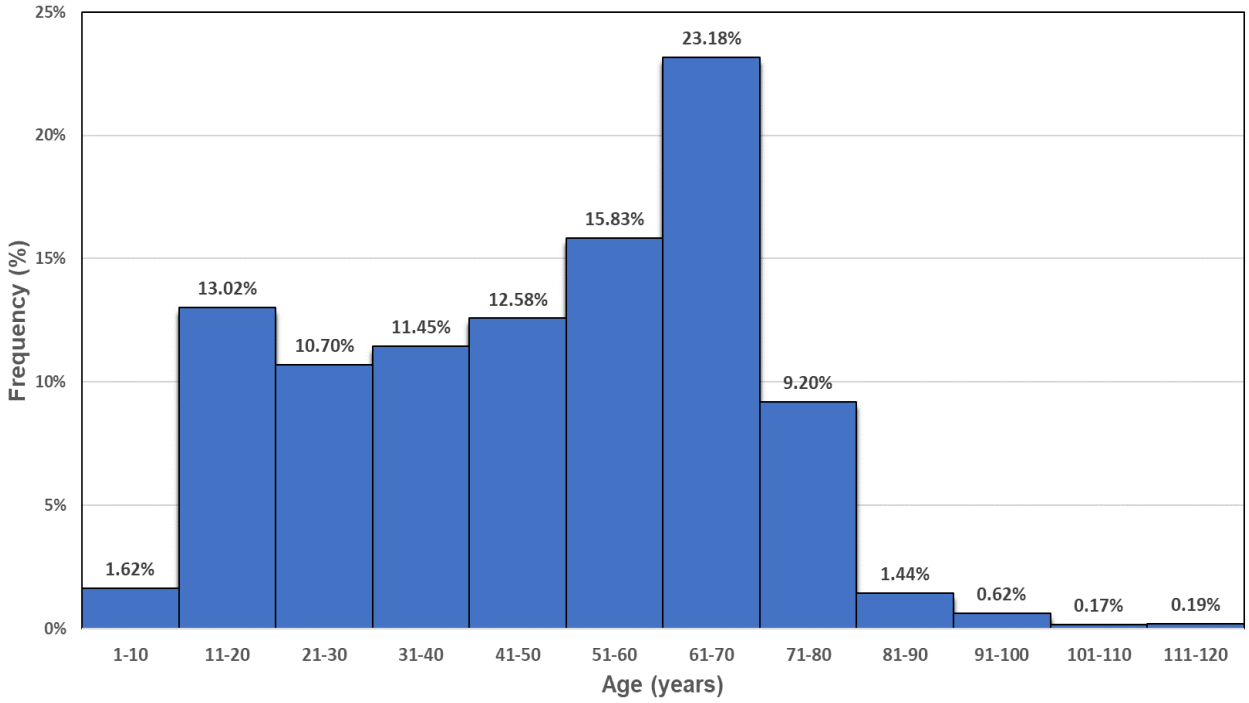

Figure 4-6 Frequency of Pipe Age

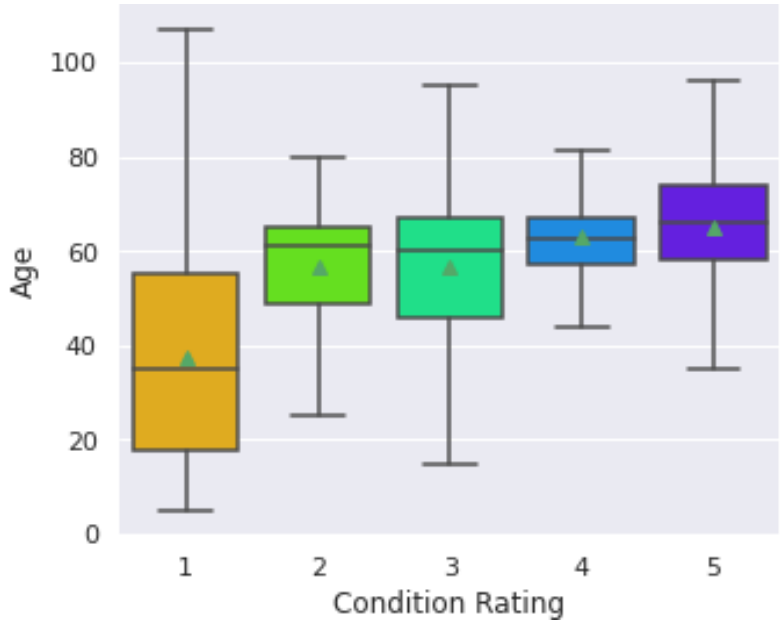

Figure 4-7 Boxplot of Age with respect to Condition Rating

4.3.2 Length

Pipe length is the manhole to manhole length of segments. The sanitary sewer dataset contains pipes ranging from 5 to 2260 feet. The pipes with a length of 200 to 300 feet have the highest frequency percentage, as shown in Figure 4-8. Only a small percentage of the sewer pipes were longer than 500 feet.

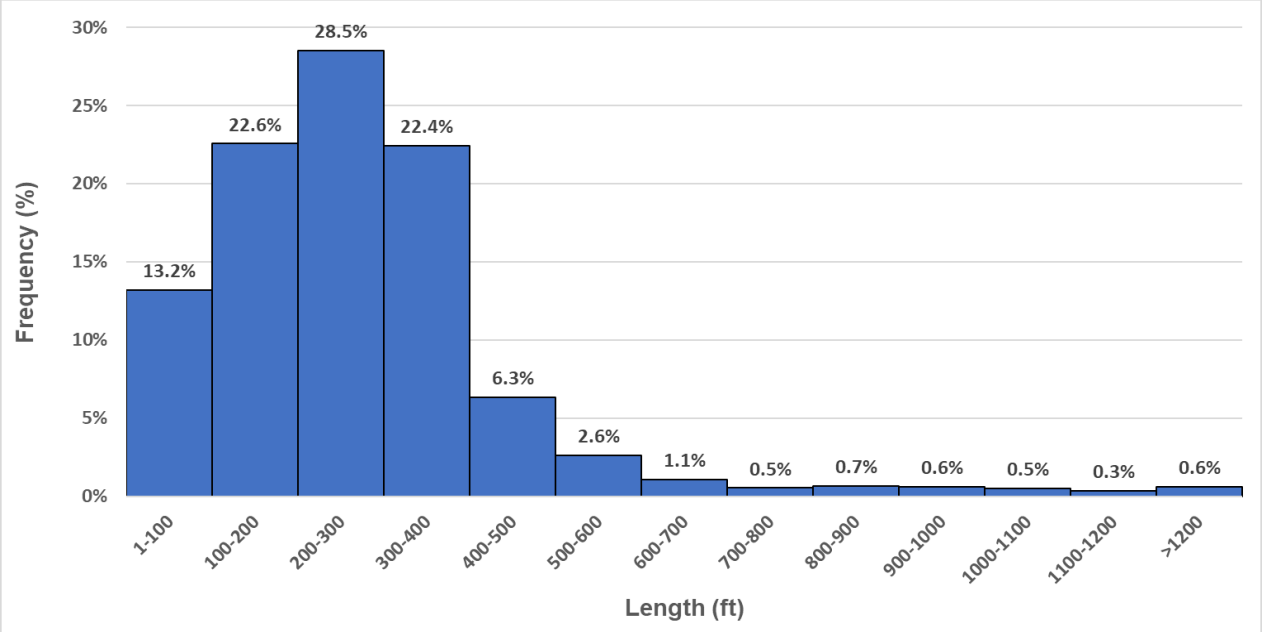

Figure 4-8 Frequency of Pipe Length

Figure 4-9 also shows that the average length of pipes with condition rating 1 is around 280 feet, and the average length of pipes with condition rating 5 is around 420 feet.

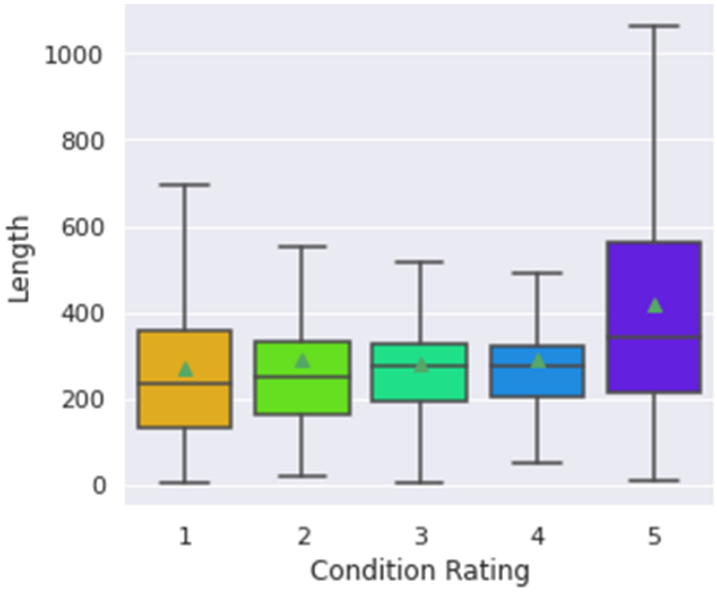

Figure 4-9 Boxplot of Length with respect to Condition Rating

4.3.3 Slope

The slope of a sewer pipe segment is calculated by dividing the difference in elevation between the upstream and downstream manholes by the length of the inspected pipe segment. It was discovered that 92 percent of the pipes have a slope of less than 2%. However, the maximum slope was found to be 55%. Figures 4-10 and 4-11 depict the slope distribution and state of pipe slope in relation to condition rating, respectively.

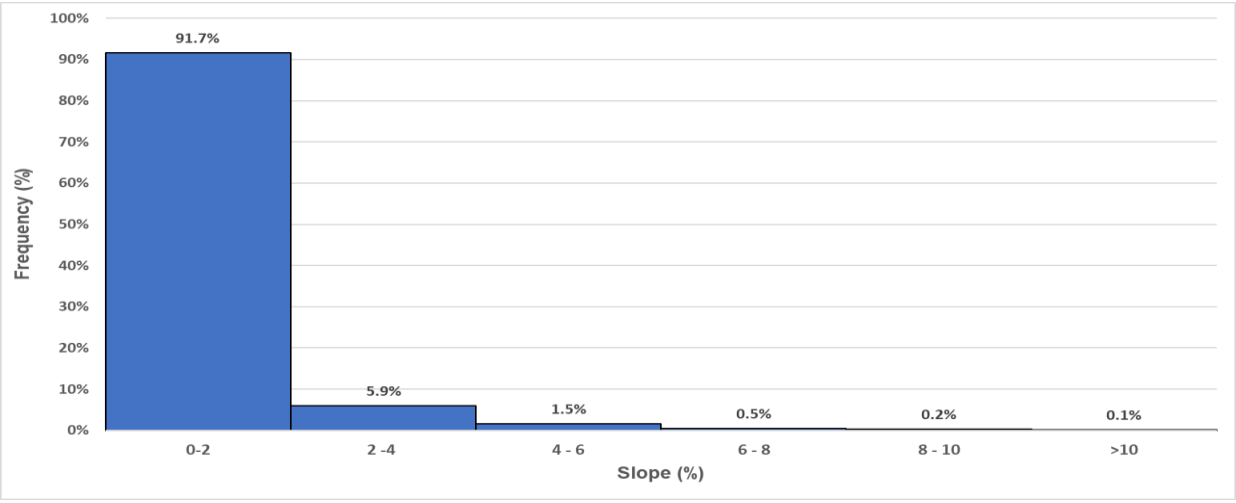

Figure 4-10 Frequency of Pipe Slope

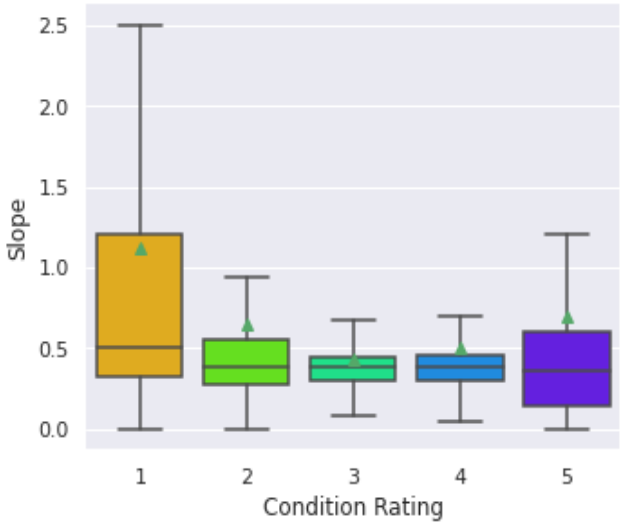

Figure 4-11 Boxplot of Slope with respect to Condition Rating

4.3.4 Diameter

Pipe diameter refers to the size of sanitary sewer pipes. The minimum and maximum sizes are 8 and 90 in., respectively. The majority of the pipes have a diameter of 8 in., as shown in Figure 4-12. Only about 13% of the pipes in the dataset have a diameter larger than 12 in.. In addition, Figure 4-13 illustrates range of pipe size concerning condition rating. It is shown that the average size of pipes in poor condition is around 18 in., and the pipes with smaller average sizes are in better condition.

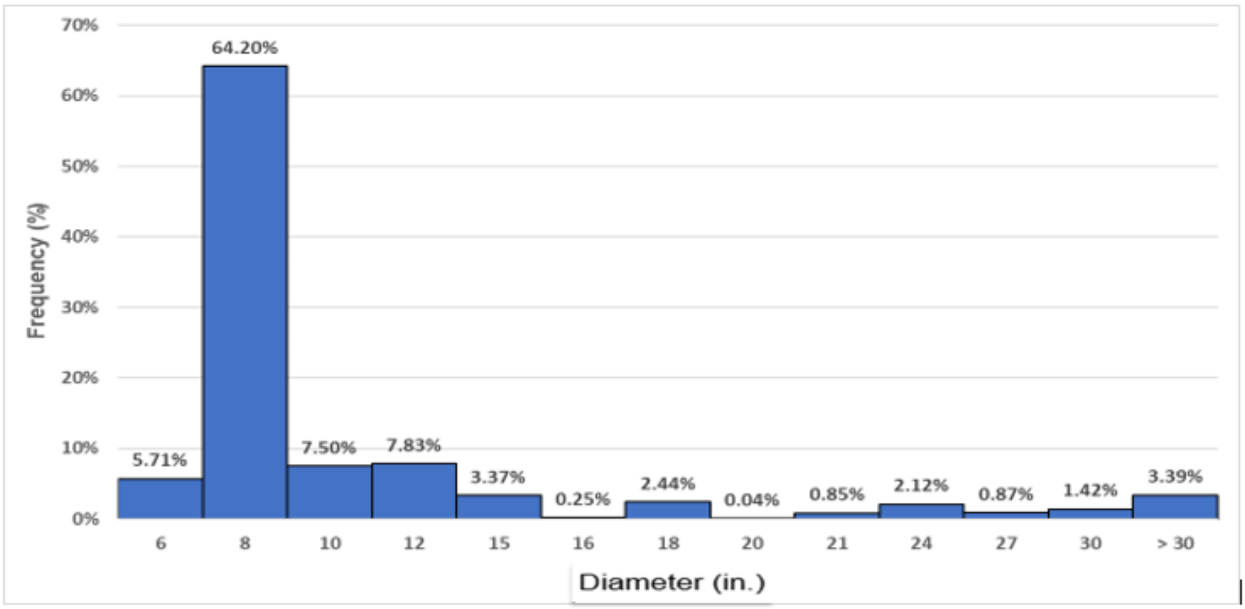

Figure 4-12 Frequency of Pipe Size

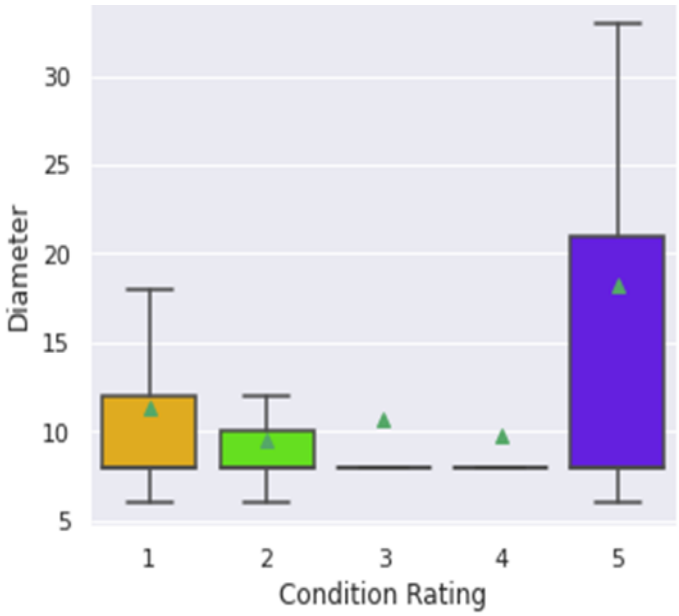

Figure 4-13 Boxplot of Diameter with respect to Condition Rating

4.3.5 Depth

The depth of a sanitary sewer pipe is measured in ft from the top of the pipe to the backfill. According to Figure 4-14, the majority of the pipes were buried between 4 and 10 ft deep, with 30.8 percent of the pipes covered by 6 to 8 ft of backfill. Just a few pipe segments are buried under depths of fewer than 4 ft and more than 12 ft. The average depth is 8.4 ft, and the maximum depth is 78 ft.

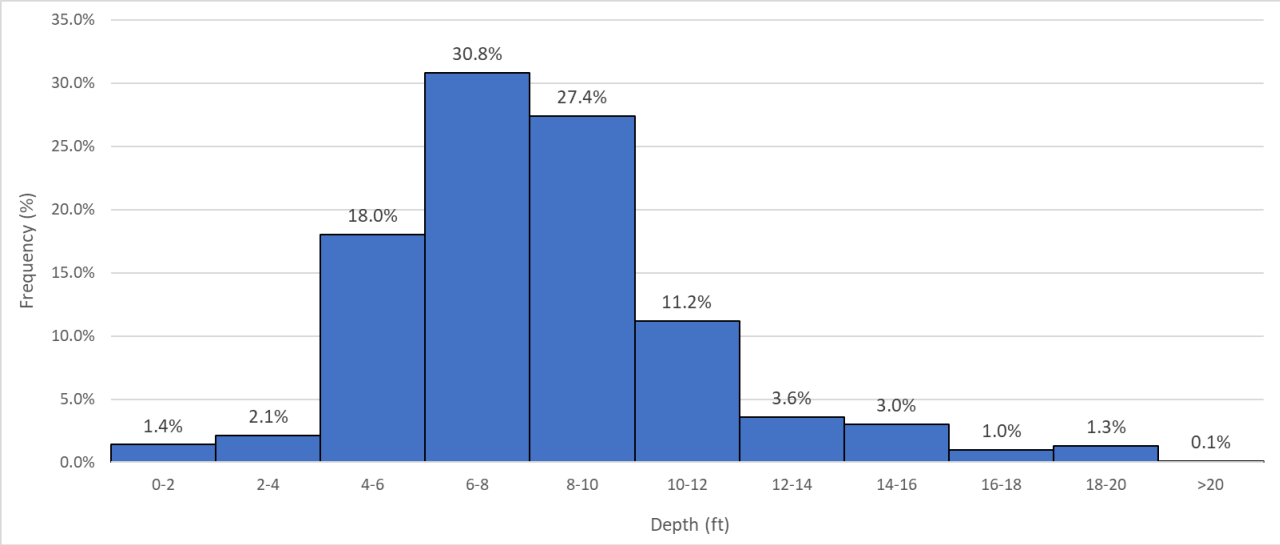

Figure 4-14 Frequency of Depth

Figure 4-15 depicts the depth range for condition ratings. It can be seen that the average depth does not differ significantly across all conditions.

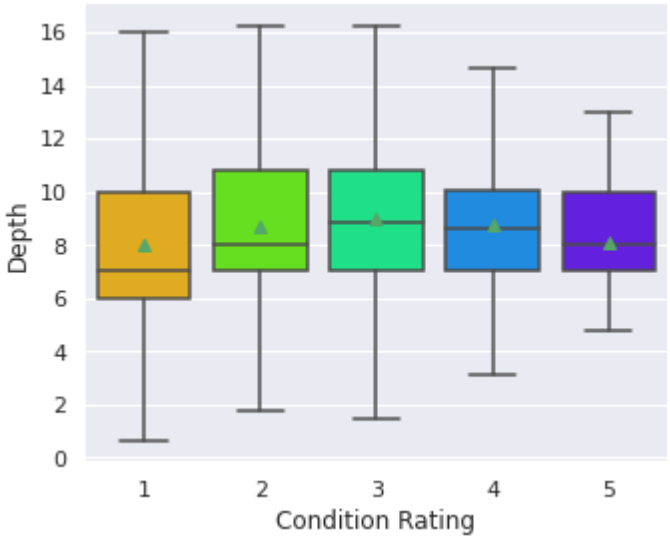

Figure 4-15 Boxplot of Depth with respect to Condition Rating

4.3.6 Soil pH

The soil pH is a numerical expression of the relative acidity or alkalinity of a soil sample. The pH range can be classified as alkaline (pH>7), natural (pH=7), or acidic (pH<7). The pH distribution in the sanitary sewer dataset is depicted in Figure 4-16. The pH values in the available dataset range from 4 to 8, with a mean of 6.87. Based on the histogram, 48 percent of soil areas have a pH of less than 7, indicating a high risk of acidity and corrosion. The average pH of each condition rating is shown in Figure 4-17.

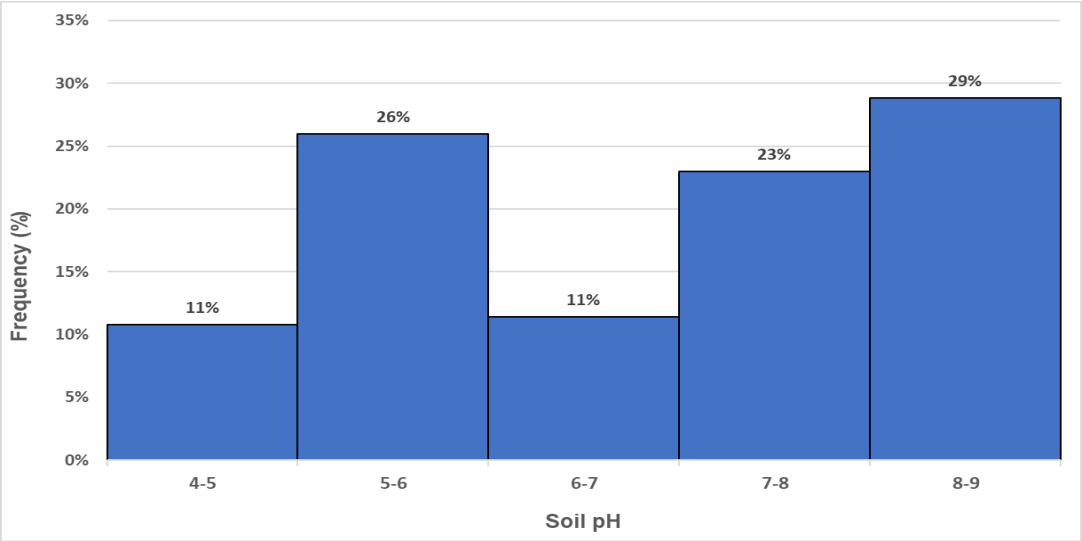

Figure 4-16 Frequency of Soil pH

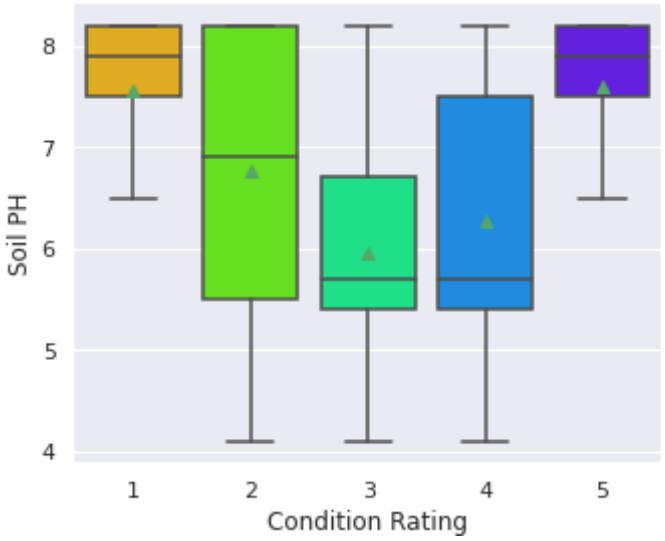

Figure 4-17 Boxplot of Soil pH with respect to Condition Rating

4.3.7 Material

The dataset involved different type of pipe material such as asbestos-cement (AC), cast iron (CAS), ductile iron pipe (DIP), High density polyethylene pipe (HDPE), prestressed concrete cylinder pipe (PCCP), polyvinyl chloride (PVC), reinforced concrete (RC), and vitrified clay pipe (VCP). In this dissertation only RC, PVC, and VCP pipes are included. Since other materials had a small portion of dataset, they are excluded to avoid any error during model development. Figure 4-18 shows vitrified clay pipes have the majority with 53% frequency. Based on Figure 4-19, most of PVC pipes are in good condition, while RC pipes have the largest share among the pipes with poor condition.

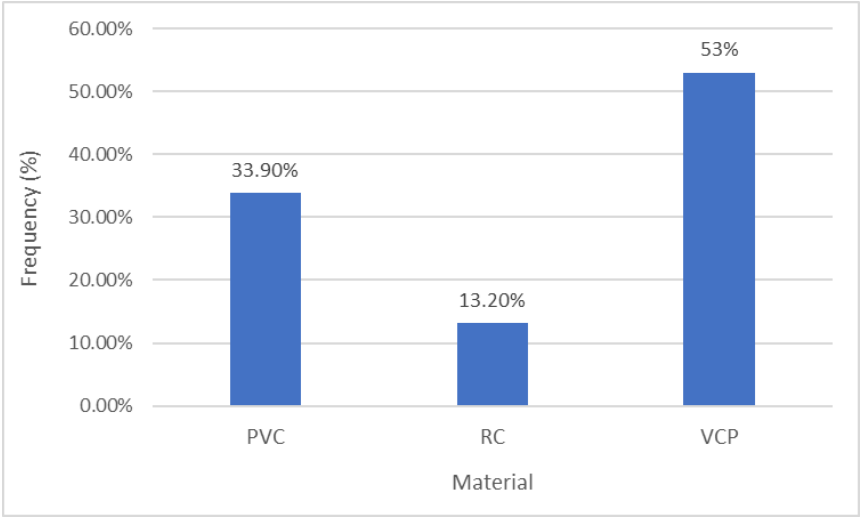

Figure 4-18 Frequency of Pipe Material

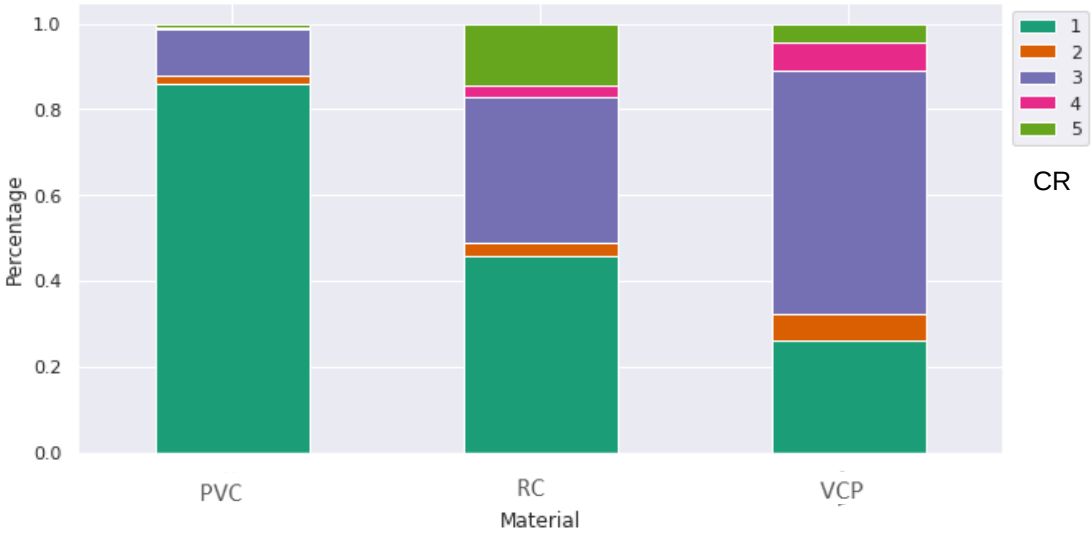

Figure 4-19 Distribution of Material with respect to Condition Rating (CR)

4.3.8 Soil Type

Figure 4-20 shows the frequency distribution of sewer pipes' soil types. There are five soil types: clay, gravel, loam, rock, and sand. It can be seen that sand has the largest frequency with 48%, and then around 36% of the sewer pipe segments are installed in locations with clay soil. Sewer pipes in the gravel soil have the lowest frequency (2%). Figure 4-21 explains condition of pipes located in different types of soils. According to Figure 3-21, the most frequent condition rating of 1 belongs to pipes surrounded by rock which is around 80%. On the other hand, around 10% of pipes surrounded by clay have a condition rating of 5, the largest among five different soil types.

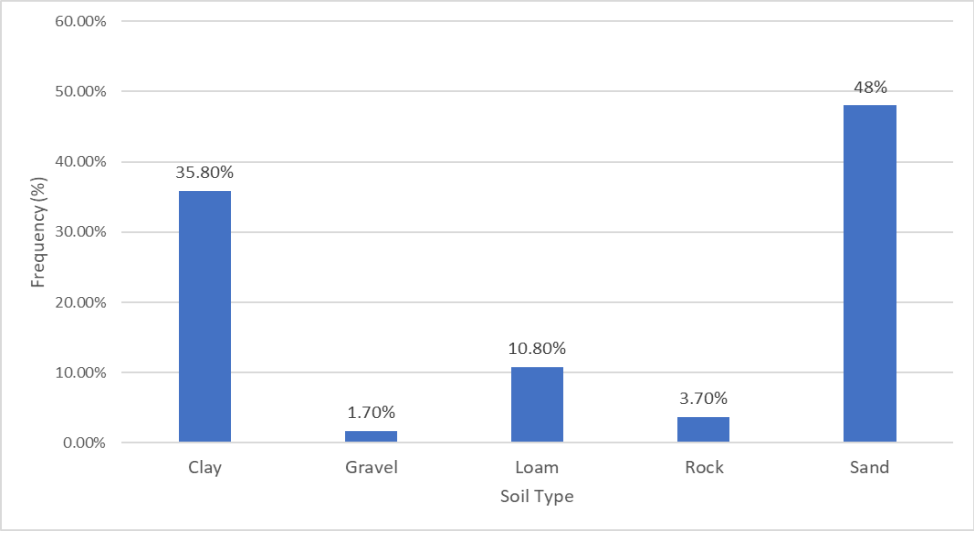

Figure 4-20 Frequency of Soil Type

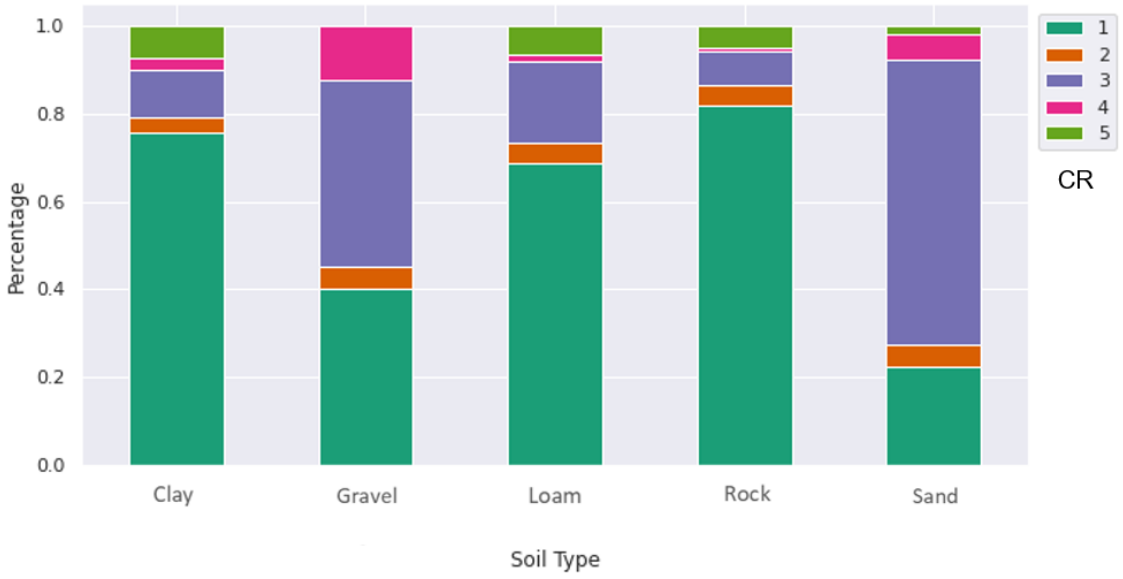

Figure 4-21 Distribution of Soil Type with respect to Condition Rating (CR)

4.3.9 Pipe Location

Figure 4-22 shows the distribution of surface conditions in the location where the pipes were buried. It reveals that sewer pipes located beneath streets were 66% of the total segments in the datasets. This was followed by sewer pipes located beneath alleys (15%), easements (11%), and Highways (8%), respectively, from the highest frequency to the least. According to the Federal Highway Administration Statistics (FHWA, 2011), highways generally have more VMT (Vehicles Miles Traveled), which shows more load on the surface. Then, the street, the alley, and the easement are in the following ranks, respectively. Furthermore, Figure 4-23 illustrates distribution of pipe locations with respect to condition rating. Based on it, segments beneath easements have the lowest percentage of pipes in poor condition.

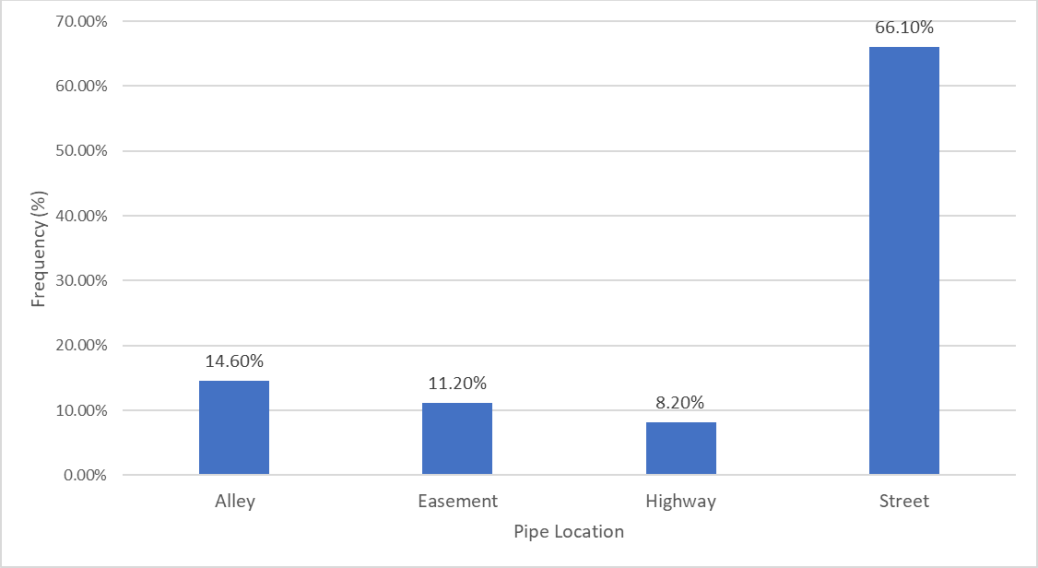

Figure 4-22 Frequency of Pipe Location

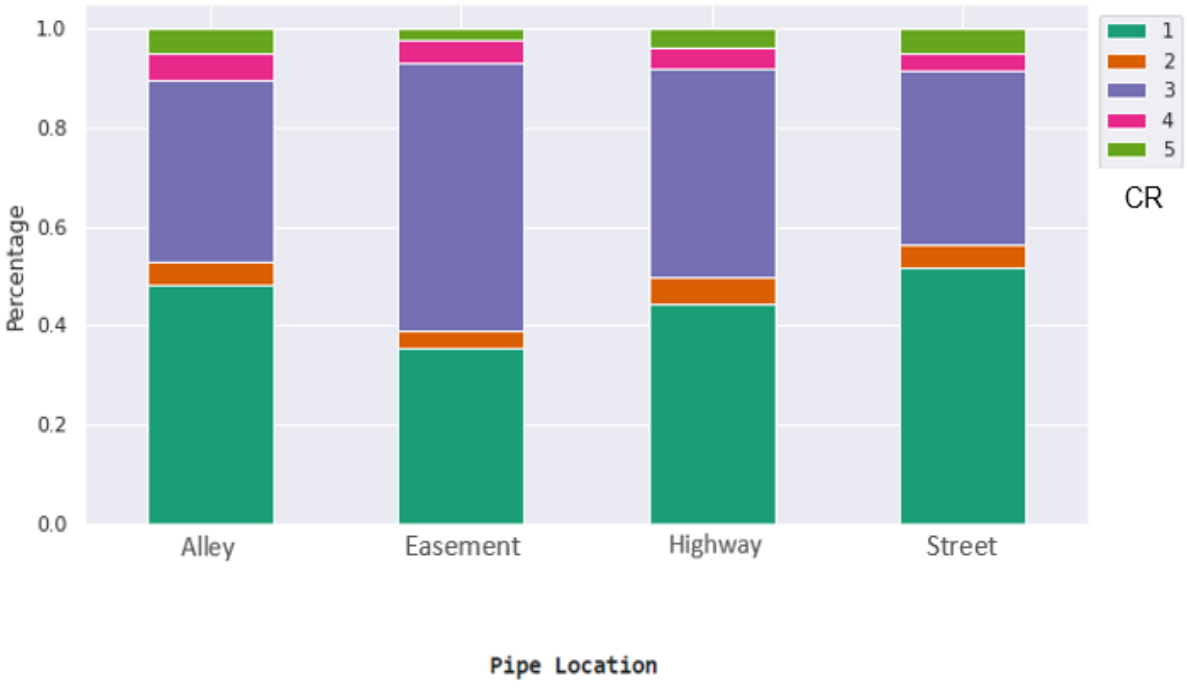

Figure 4-23 Distribution Pipe Location with respect to Condition Rating (CR)

4.3.10 Condition Rating

The condition rating (dependent variable) of sewer pipes was predicted using nine different independent variables in this study. As previously stated, the scores range from 1 to 5, with 1 indicating structurally sound pipes and 5 indicating pipes on the edge of failure. Figure 4-24 illustrates the distribution

of pipe conditions in the dataset. Pipes with excellent conditions (rating of 1) have the highest frequency with 49%, followed by pipes with a rating of 3 with 38%. The lowest frequency belongs to pipes with a rating of 4. Interestingly, only 4.4% of pipe segments have a rating of 2. It can be seen that around 10% of pipes are in poor condition (ratings 4 and 5). It should also be noted that 75% of the pipes are less than 65 years old and around 90% are in fair condition (ratings 1,2 and 3).

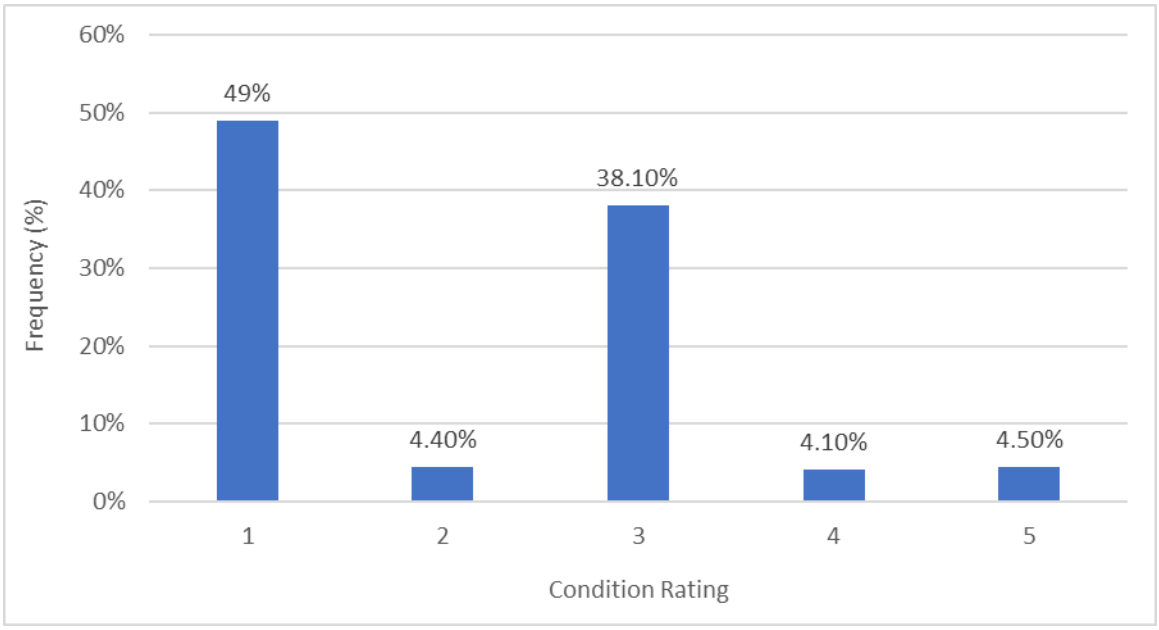

Figure 4-24 Frequency of Pipe Condition Rating

4.4 Descriptive Statistics

Descriptive statistics provide a summary of quantitative analysis for the dataset's numerical variables. It shows a summary of the data sample and variable characteristics in the sewer dataset. The descriptive statistics of numerical variables in this study are presented in Table 4-3. For example, it shows the minimum, maximum, and average age of pipes in the available dataset. Also, it demonstrates that 25% of pipes are less than 30, 50% less than 52, and 75% less than 65 years old. Std (Standard deviation) is a measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean of the set, while a high standard deviation indicates that the values are spread out over a wider range.

Table 4-3 Descriptive Statistics of Numerical Variables

	Mean	Min	25%	50%	75%	Max	Std
Age	47.78	1.00	30.00	52.00	65.00	121.00	20.72
Diameter	11.26	6.00	8.00	8.00	10.00	90.00	9.18
Slope	0.79	0.00	0.30	0.40	0.67	55.66	1.57
Depth	8.44	0.09	7.00	8.00	10.00	78.00	3.31
Length	282.06	5.47	159.99	261.37	337.53	2261.24	198.06
Soil pH	6.87	4.10	5.70	7.50	8.20	8.20	1.31

4.5 Correlation Analysis

Correlation analysis is a statistical method for determining the degree of relationship between two different variables. This relationship can range from strong to weak, and there are times when there is no relationship between two variables. A strong relationship means that the value of one variable can be predicted based on the value of the other variable. In contrast, when the variables' relationships are weak, they cannot be predicted well. The correlation coefficient between variables can be positive or negative and can only be between -1.00 and +1.00. Incorporating highly correlated independent variables into a model may result in a multicollinearity problem that affects the model's outcomes. Developing a model with highly correlated independent variables is not advised. Most of the variables in the available dataset were not normally distributed in the model. As a result, spearman's rank correlation was used to examine the relationship between the variables. Spearman's rank correlation can be used to describe the relationship between nonlinearly related variables. This method makes no assumptions about the distribution of the variables in the model, unlike Pearson's method, which assumes the distribution of two variables is normal and can only describe a linear relationship between two variables.

According to the result of Spearman rank analysis shown on Table 4-4, the highest correlation coefficient is between age and condition rating (+0.48). There is no strong correlation between independent variables which means none of them needs to be removed from the model to avoid multicollinearity.

Table 4-4 Correlation Analysis

	Age	Diameter	Slope	Depth	Length	Soil pH	Condition Rating
Age	1	-0.018	-.221**	.114**	.105**	-.213**	.482**
Diameter	-0.018	1	-.322**	.049**	.101**	.131**	-.076**
Slope	-.221**	-.322**	1	-.384**	-.172**	.313**	-.265**
Depth	.114**	.049**	-.384**	1	-0.006	-.265**	.164**
Length	.105**	.101**	-.172**	-0.006	1	0.014	.119**
Soil PH	-.213**	.131**	.313**	-.265**	0.014	1	-.442**
Condition Rating	.482**	-.076**	-.265**	.164**	.119**	-.442**	1

** Correlation is significant at the 0.01 level (2-tailed).

4.6 Chapter Summary

This chapter discussed data collection, data preparation, and data processing. Preliminary and explanatory data analysis, descriptive statistics, and correlation analysis were presented too. Model development using the prepared dataset is discussed in the next chapter.

Chapter 5 Model Development Results and Comparison

5.1 Introduction

This chapter explains the development process of models whose concepts were discussed in chapter 3. The strategy to avoid overfitting and resolve imbalanced data problems is presented. The results of Binary and Multinomial Logistic Regressions, k-Nearest Neighbors modeling, and tree-based models, including Decision Tree, Random Forest, AdaBoost, Gradient Boosting Tree, and XGBoost, are thoroughly discussed and compared. Also, the effect of different factors extracted from the most accurate model on the deterioration of sanitary sewer pipes is presented.

5.2 SPSS and Python

The software used to develop the statistical models (Binary Logistic Regressions) was IBM SPSS Statistics packages (SPSS 25). SPSS (Statistical Package for the Social Sciences) is a software package used for the analysis of statistical data. Although the name of SPSS reflects its original use in the field of social sciences, its use has since expanded into other data markets.

In this study, artificial intelligence models were developed using Python, one of the most popular programming languages in the data science industry. Python was utilized since it is open-source and has a wide variety of free add-on libraries. Some of the libraries used in this study are shown in the Table 5-1.

Table 5-1 Python Libraries Used in the Study

	Name of the Library	Description or Functions of the Library
1	Pandas	To open spreadsheet files and manipulate numerical tables
2	Numpy	Used to perform a wide variety of mathematical operations on arrays
3	Scikit learn	It provides a selection of efficient tools for modeling including classification, regression, clustering, and dimensionality reduction via a consistence interface in Python
4	Seaborn	It is a data visualization library
5	Matplotlib	It is a cross-platform for graphical plotting

5.3 Cross Validation

The most common validation method used in any prediction issue is cross-validation. The fundamental idea underlying cross-validation is that a portion of the input dataset is left out during model training and used during model testing. The entire dataset will be used for training and testing the model, a crucial component of the cross-validation technique (Loganathan et al., 2022; Loganathan, 2021). The main benefit of employing k-fold cross-validation is that it would prevent overfitting and allow for the inclusion of samples from all classes during model training.

The dataset is divided into K equal-sized files, which is the main concept. For instance, if there are 200 data points and ten folds (k=10), the dataset is divided into ten equal portions, and there will be 20 data points in each folder. Nine of the ten parts will be used to train the model, and one part will be used to test the trained model. In this study, 5-fold cross-validation was used for all models, as shown in Figure 5-1, because increasing the folds would reduce the number of data points in each portion. The dataset was divided into five folders using the Sklearn package of Python. Five separate learning experiments were carried out in 5-Fold cross-validation. One folder was chosen for each iteration for testing purposes, and the training set was constructed by combining the remaining four folders. This process performed five distinct iterations, and the model's output was then the average value (Malek Mohammadi, 2019).

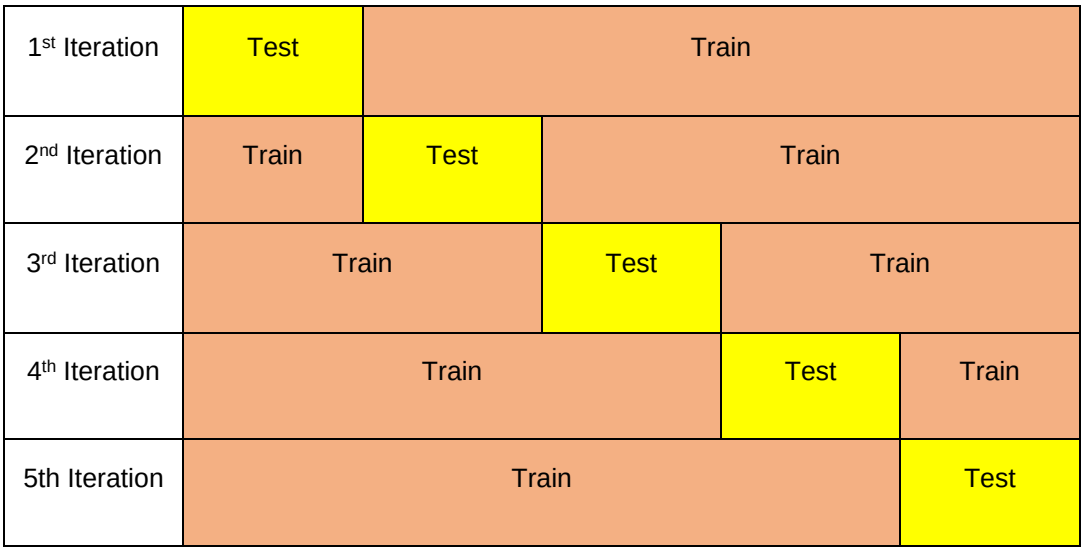

Figure 5-1 Five-Fold Cross Validation

5.4 Resampling

In the available dataset used for this study, condition ratings of 2, 4, and 5 have much fewer instances than ratings of 1 and 3, which is referred to as a dataset imbalance. When imbalanced data is employed in classification methods, trained models may perform poorly (Loganathan, 2021; Tanha et al., 2020; Yijing et al., 2016). The minority class must be given more consideration in prediction modeling for sewer pipes since the consequences of misclassifying the minority class would be significantly worse than for the other classes. Because it would be worse to misclassify the PACP scores of the 4 and 5 classes as 1, they are given more importance in this study. During model development, it was found that the performance of models (assessed by evaluation metrics) was not much good due to an imbalanced dataset. Researchers and data scientists use a variety of treatment approaches to improve outcomes from an imbalanced dataset. There are different types of resampling, including over-sampling, under-sampling, and hybrid methods (Ghorbani and Ghousi, 2020):

- Over-Sampling Method: By replicating existing minority class samples or creating new ones, oversampling increases the weight of the minority class. There are various over-sampling techniques, and it's important to note that the over-sampling strategy is typically used more frequently than other strategies.
- Under-Sampling Method: Under-sampling is one of the simplest methods to address the issue of unbalanced data. To balance the class with the minority class, this method under-samples the majority class. When there is enough data obtained, the under-sampling approach is used.
- Hybrid Method: There are several benefits and drawbacks to both over-sampling and under-sampling. The advantages and disadvantages of each approach can be obtained by combining these two techniques.

Various methods of each resampling approach containing their description are shown in Table 5-2.

Table 5-2 Different Resampling Methods

(Ghorbani and Ghousi, 2020; Taneja et al., 2019)

Approach	Method	Description
Over-Sampling	SMOTE*	Balances class distribution by synthetically generating new minority class instances along directions from existing minority class instances towards their nearest neighbors.
	Borderline-SMOTE	Generates the synthetic sample along the borderline of minority and majority classes.
	SVM-SMOTE	Focuses on generating new minority class instances near borderlines with SVM.
	ADASYN*	Balances minority instances using regular SMOTE algorithm with adding random small values to the points.
Under-Sampling	Down-Sampling	Majority class instance is reduced to the size of minority class by eliminating randomly some majority class instances.
Hybrid	SMOTE-ENN	This method is one of the well-known methods that combines the SMOTE as over-sampling model and ENN (Edited Nearest Neighbors) as an under-sampling model to improve the results.
	SMOTE-Tomek	This method is another common hybrid method that connects the SMOTE as an over-sampling model to Tomek links as an under-sampling model to enhance the results.

*SMOTE: Synthetic Minority Oversampling Technique

* ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced

Various over-resampling and hybrid techniques were used in this study to resolve the problem of imbalanced datasets such as SMOTE, SVM-SMOTE, Borderline-SMOTE, SMOTE-ENN, SMOTE-Tomek. The under-sampling approach was not used because we had trouble with a low number of data. Table 5-3 lists all of the resampling techniques utilized in this study along with the values for each of their key parameters. These settings yield the best results.

Table 5-3 Methods of resampling with associated parameter settings

Method	Parameters
SMOTE	K_Neighbors = 7
Borderline-SMOTE	K_Neighbors = 7, M_Neighbors = 10
SVM-SMOTE	K_Neighbors = 7, M_Neighbors = 10
SMOTE-ENN	K (SMOTE)= 7, K (ENN) = 3
SMOTE-Tomek	K (SMOTE)= 7

Based on the F1-score of machine learning methods using resampled datasets, it was found that SVM-SMOTE has a better effect. So, different models described in the following sections of the current chapter are based on resampled data by the SVM-SMOTE method. Section 5.4.3 explains this method's details.

5.4.1 SMOTE

Regular SMOTE is a statistical method that produces new instances to enhance the number of minority samples in the dataset. This algorithm creates new samples by combining the target case's features with its neighbors' features after sampling the feature space for each target class and its closest neighbors. The new samples do not replicate the minority samples from the past (Chawla et al., 2002). Figure 5-2 depicts SMOTE algorithm. SMOTE follows a simple approach:

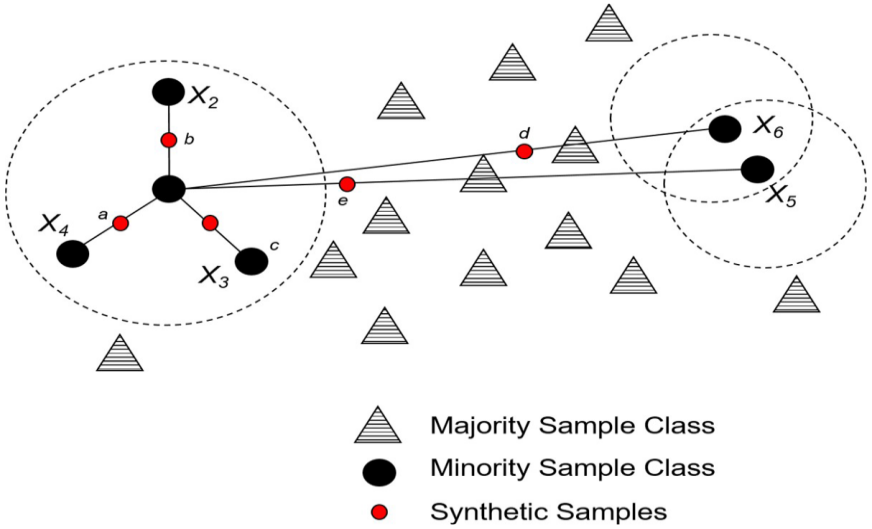

Figure 5-2 SMOTE Algorithm with k=5

(Teh et al., 2020)

1. Select a sample, let's call it ***O*** (for Origin), from the minority class randomly.

2. Find the K-Nearest Neighbors of O that belong to the same class.

3. Connect O to each of these neighbors using a straight line.

4. Select a scaling factor '***z'*** in the range [0,1] randomly.

5. For each new connection, place a new point on the line (z*100)% away from O. These will be our synthetic samples.

6. Repeat this process until you get the desired number of synthetic samples.

5.4.2 Borderline-SMOTE

It is an adaptation of the SMOTE. Borderline-SMOTE only creates synthetic data along the decision boundary between the two classes, unlike SMOTE, which generates them arbitrarily between the two data. In the training process, Borderline-SMOTE algorithms try to learn the borderline of each class, where these borderline cases and the ones adjacent are more likely to be misclassified than the ones distant from the borderline. In Borderline-SMOTE, all minority instances are categorized into three groups: NOISE, DANGER, and SAFE by calculating the K-nearest neighbors of each minority instance and the numbers of the majority samples (m) that were discovered in K nearest neighbors of this instance (Zheng, 2020; Elnahas et al., 2021). The three regions are defined according to Table 5-4.

Table 5-4 Borderline-SMOTE Regions

(Elnahas et al., 2021)

Region	Definition
Noise	m=k
Safe	0 ≤ m ≤ k/2
Danger	k/2 ≤ m ≤ k

After all instances of the minority class are categorized, synthetic instances are then created along the line between DANGER instances and their nearest neighbors Figure 5-3 shows different groups of minority class samples and created synthetic samples by Borderline-SMOTE algorithm.

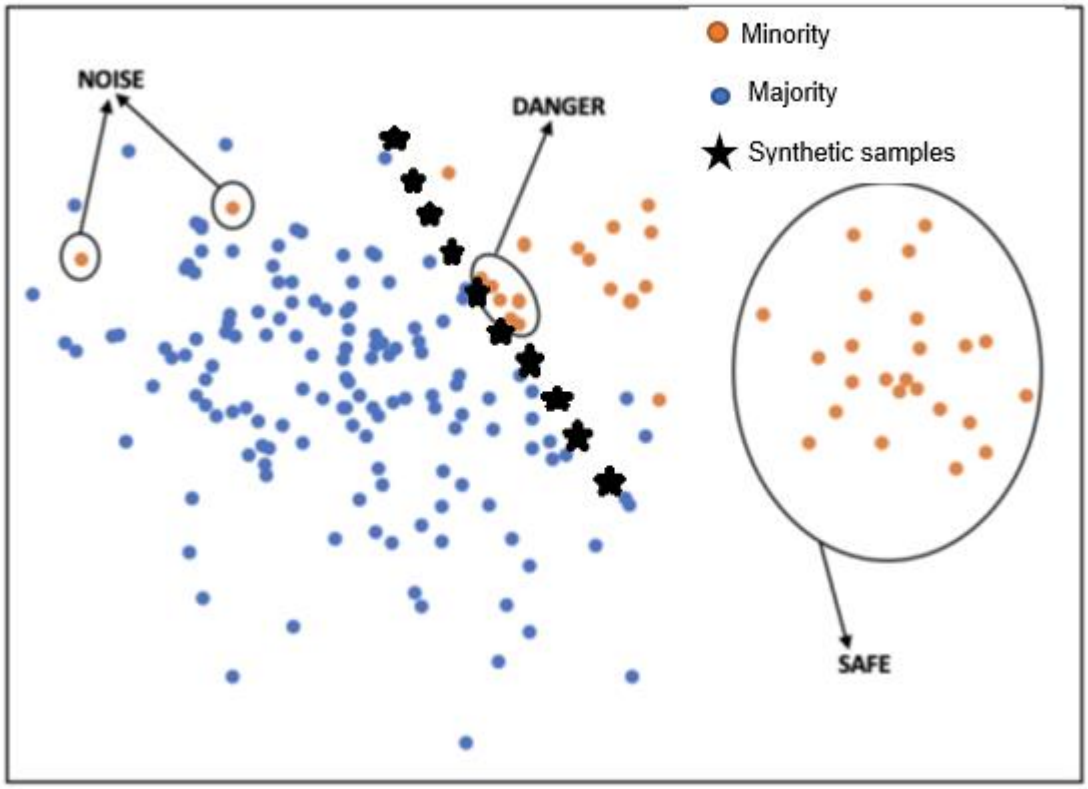

Figure 5-3 Borderline-SMOTE Algorithm

(Zheng, 2020)

5.4.3 SVM-SMOTE

Another variation of Borderline-SMOTE is Borderline-SMOTE SVM, or we could just call it SVM-SMOTE. In the SVM-SMOTE, the borderline area is approximated by the support vectors after training SVMs classifier on the original training set. Instead of utilizing the k-nearest neighbor algorithm to identify 79Borderline-SMOTE that uses the support vector machine (SVM) methodology (Zheng, 2020). Synthetic data will be randomly created along the lines joining each minority class support vector with a number of its nearest neighbors. What special about Borderline-SMOTE SVM compared to the Borderline-SMOTE is that more data are synthesized away from the region of class overlap. It focuses more on where the data is separated. The SVM-SMOTE focuses on developing new minority class samples close to class

boundaries while utilizing the SVM model to establish boundaries. By using current minority class instances as a guide to their closest neighbors, this approach creates new minority class instances (Ghorbani and Ghousi, 2020). So, it helps establish boundaries between classes while generating new instances simultaneously (Taneja et al., 2019).

The resampled data is used to develop all AI models. It has a quantity of 7939 data points which 80% of that was used to train the models (6351 data points), and 20% was used for testing the model (1588 data points). It means that the sum of numbers contained in the confusion matrix of each model should be equal to 1588 because the confusion matrixes shown in the following sections are results of model testing.

5.5 Logistic Regression

5.5.1 Binary

At the beginning of the model development process, to investigate the effect of different parameters on the condition of sewer pipes in a simple way, the state of sanitary sewer pipes in the original dataset was classified into two groups of 0 as good and 1 as poor condition, a binary classification. Therefore, the pipes with condition ratings of 1,2, and 3 were classified as group 0 and pipes with condition ratings of 4 and 5 as group 1. Figure 5-4 shows the statistics of sewer pipes in the new category. As shown in figure 5-4, the recoded dataset includes 90% and 10% pipes with condition levels 0 and 1, respectively.

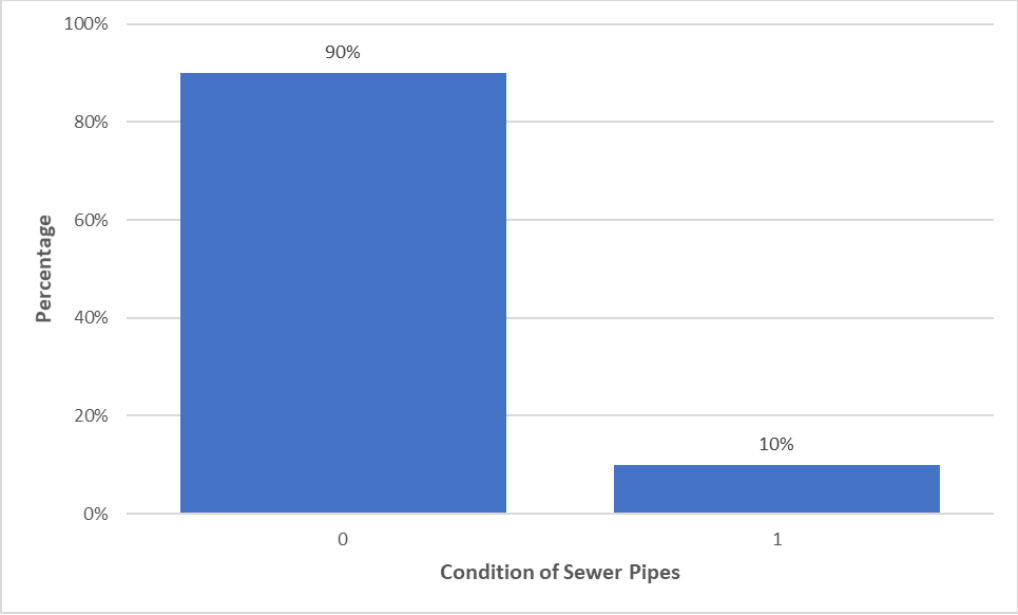

Figure 5-4 Frequency of Sewer Pipes Conditions in a Binary Classification

5.5.1.1 Training the Model

Based on the characteristics of dependent variable which has only two values, one regression equation is generated to estimate the condition of each pipe segments as shown in Equation 5.1. the SPSS software was used to develop the binary model.

$$ln\left(\frac{P(C=1)}{1-P(C=1)}\right) = \quad \text{Eq. 5.1}$$

$$\alpha + \beta_1 \times Age + \beta_2 \times Diameter + \beta_3 \times Depth + \beta_4 \times Slope + \beta_5 \times Length + \beta_6 \times pH$$
$$+ \beta_7 \times D_{Material=PVC} + \beta_8 \times D_{Material=VCP} + \beta_9 \times D_{Material=RC} + \beta_{10} \times D_{Soil=S}$$
$$+ \beta_{11} \times D_{Soil=C} + \beta_{12} \times D_{Soil=G} + \beta_{13} \times D_{Soil=L} + \beta_{14} \times D_{Soil=R} +$$
$$\beta_{15} \times D_{Road=St} + \beta_{16} \times D_{Road=H} + \beta_{17} \times D_{Road=E} + \beta_{18} \times D_{Road=A}$$

where α is intercept, $\beta_1, \beta_2, \ldots, \beta_{18}$ are regression coefficients, and D_i is dummy variable to assign different values to categorical independent variables.

In logistic regression, if the dependent variable has N categories, one of them is chosen as the reference category. In this dissertation, condition level 0 was chosen as the reference category for the development of binary logistic regression. The binary logistic regression was trained by SPSS using 80% of the data. Model parameters were estimated using Maximum Likelihood Estimation (MLE). Wald test and P-test were used to determine the variables' significance, with a 95% confidence level. The factors with the highest ability to predict the state of sanitary sewer pipes were found using a backward stepwise variable selection. In this method, the variables with enough predictive power remain in the model, and idle variables are removed stepwise. Backward stepwise selection started with a complete model and took into account all 9 independent variables; those that had the least impact on the model were then left out. The variables with the highest p-score were those being considered for model removal. Table 5-5 provides parameter estimates for the various sanitary sewer pipe conditions.

Table 5-5 Parameter Estimates in Binary Logistic Regression for Condition Level 1

Variable	Coefficient (β)	Standard Error	Wald	Sig. (P-Value)	Exp(β) (Expected Value)
Age	0.043	0.004	113.23	0	1.044
Diameter	0.007	0.005	1.71	0.191	1.007
Slope	-0.002	0.026	0.007	0.933	0.998
Depth	-0.005	0.017	0.105	0.745	0.995
Length	0.001	0	19.781	0	1.001
Soil pH	0.065	0.094	0.485	0.486	1.067
Material (VCP) (Reference)	0	-	-	0	-
Material (PVC)	-0.974	0.275	12.586	0	0.378
Material (RC)	-0.087	0.15	0.337	0.562	0.916
Soil Type (Sand) (Reference)	0	-	-	0	-
Soil Type (Clay)	0.442	0.246	3.236	0.072	1.556
Soil Type (Gravel)	0.628	0.37	2.882	0.09	1.874
Soil Type (Loam)	-0.007	0.3	0.001	0.981	0.993
Soil Type (Rock)	-0.141	0.408	0.12	0.729	0.868
Road Type (Street) (Reference)	0	-	-	0.144	-
Road Type (Alley)	0.215	0.139	2.368	0.124	1.239
Road Type (Easement)	-0.241	0.183	1.737	0.188	0.786
Road Type (Highway)	-0.124	0.198	0.396	0.529	0.883
Constant	-5.565	0.626	79.091	0	0.004

Table 5-6 illustrates the result of the backward stepwise method. Utilizing this method, the least important variables were removed from the full model. Remained parameters are the most important ones. As can be seen in Table 5-6, age, length, material, and soil type had the least P-values and consequently remained in the model.

Table 5-6 Parameter Estimates in Binary Logistic Regression (Backward Stepwise)

Variable	Coefficient (β)	Standard Error	Wald	P Value
Age	0.043	0.004	112.913	0
Length	0.001	0	27.094	0
Material (VCP) (Reference)	0	-	12.893	0
Material (PVC)	-0.976	0.273	12.761	0
Material (RC)	-0.12	0.143	0.02	0.887
Soil Type (Sand) (Reference)	0	-	25.911	0
Soil Type (Clay)	0.609	0.133	20.804	0
Soil Type (Gravel)	0.667	0.352	3.577	0.059
Soil Type (Loam)	0.206	0.203	1.032	0.31
Soil Type (Rock)	0.123	0.338	0.005	0.945
Constant	-5.197	0.277	351.058	0

5.5.1.2 Results and Discussions

The developed binary logistic regression provided one equation to predict the condition ratings of sewer pipes based on available dataset. As significant variables were found, the model created the equation utilizing only those variables. The results of binary logistic regression are shown in Equation 5.2.

$$g(x) = ln\left(\frac{P(C=1)}{1-P(C=1)}\right) = \quad \text{Eq. 5.2}$$

$$-5.197 + 0.043\ x\ Age + 0.001\ x\ Length$$
$$- 0.976 \times D_{PVC} - 0.12\ X\ D_{RC}$$
$$+ 0.609\ x\ D_{clay} + 0.667\ x\ D_{gravel} + 0.206\ x\ D_{loam} + 0.123\ x\ D_{rock}$$

Once the odds ratio (g(x)) was calculated, the probability of pipes being in poor or good condition can be estimated by using Equation 5.3 and 5.4, respectively.

$$P(C = 1) = \frac{1}{1+e^{-g(x)}} \quad \text{Eq. 5.3}$$

$$P(C = 0) = 1 - P(C = 1) \quad \text{Eq. 5.4}$$

Table 5-7 shows the classification table for the developed Binary Logistic Regression Model. It is an evaluation of the effectiveness of this model. 20% of data has been used to test the model and create this table.

Table 5-7 Classification Table for Binary Logistic Regression

Observed	Predicted		Percent Correct Predicted
	0	1	
0	804	61	93%
1	45	51	53%
Overall Percentage			88%

Overall, binary logistic regression was able to accurately predict 88% of the pipe conditions, according to the classification table. The prediction accuracy for the pipes at condition levels 0 and 1 were 93% and 53%, respectively. Based on the table above and using equations described in section 3.5.1, it is possible to calculate True Positive, True Negative, False Positive, and False Negative rates. Table 5-8 and depicts these numerical values.

Table 5-8 Binary Logistic Regression Model Performance

Rates	Values
True positive rate (TPR)	93%
True negative rate (TNR)	53%
False positive rate (FPR)	47%
False negative rate (FNR)	7%

The performance of the model is shown by the area under the ROC curve, where ideal models have an area close to 1, and weaker models have an area near to 0.5. The model is considered satisfactory if the area under the ROC curve is bigger than 0.7 (Malek Mohammadi, 2019; Hosmer et al., 2013). The ROC curve for developed Binary Logistic Regression is shown in Figure 5-5. The area under ROC curve is 0.73 which shows binary logistic regression have had acceptable result.

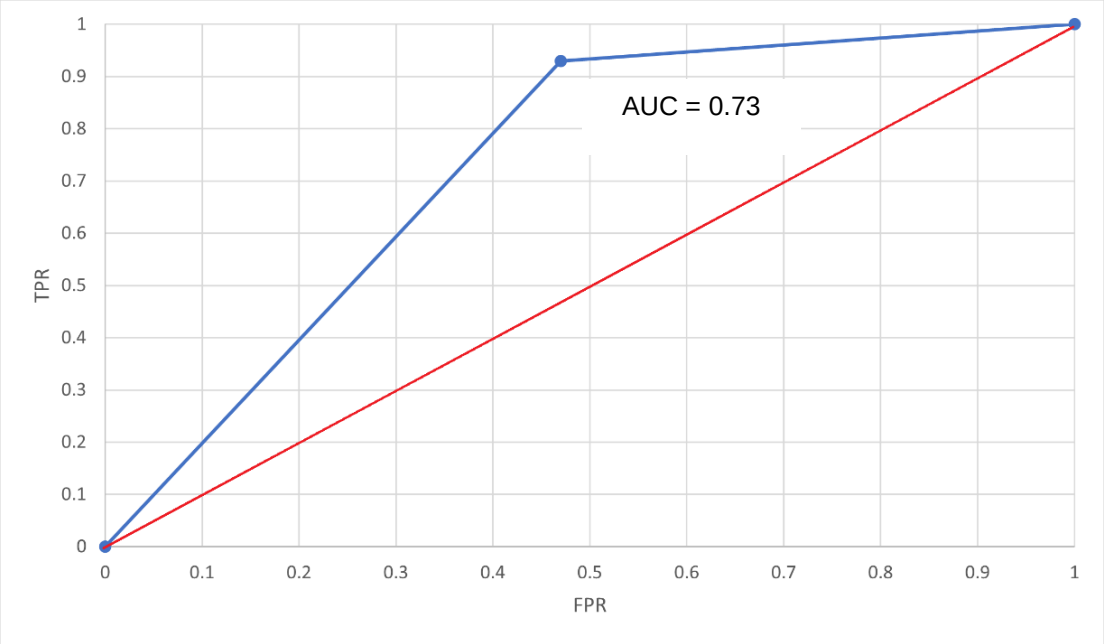

Figure 5-5 ROC Curve for Binary Logistic Regression Model

It is possible to create a visual representation of the likelihood that pipes are in poor or good condition using binary logistic regression results. The deterioration curves were created in this study to demonstrate how the state of sewer pipes deteriorates over time while accounting for important variables. According to backward selection method used in development of the binary logistic regression, age, length, material, and soil type were significant parameters in pipes deterioration. The mean values of the independent numerical variables were used to create the degradation curves. The average age of pipes in available dataset was 48 years old and the average length of them was 283 ft. Using equations 5.2 and 5.3, the probability of sewer pipes being in poor conditions was visualized and is shown in Figures 5-6 to 5-12.

The deterioration curve of sewer pipes for three different pipe materials buried in the sand and clay, are shown in Figures 5-6 and 5-7. Only pipes buried in sand and clay were investigated since they are the majority of soil types in the dataset.

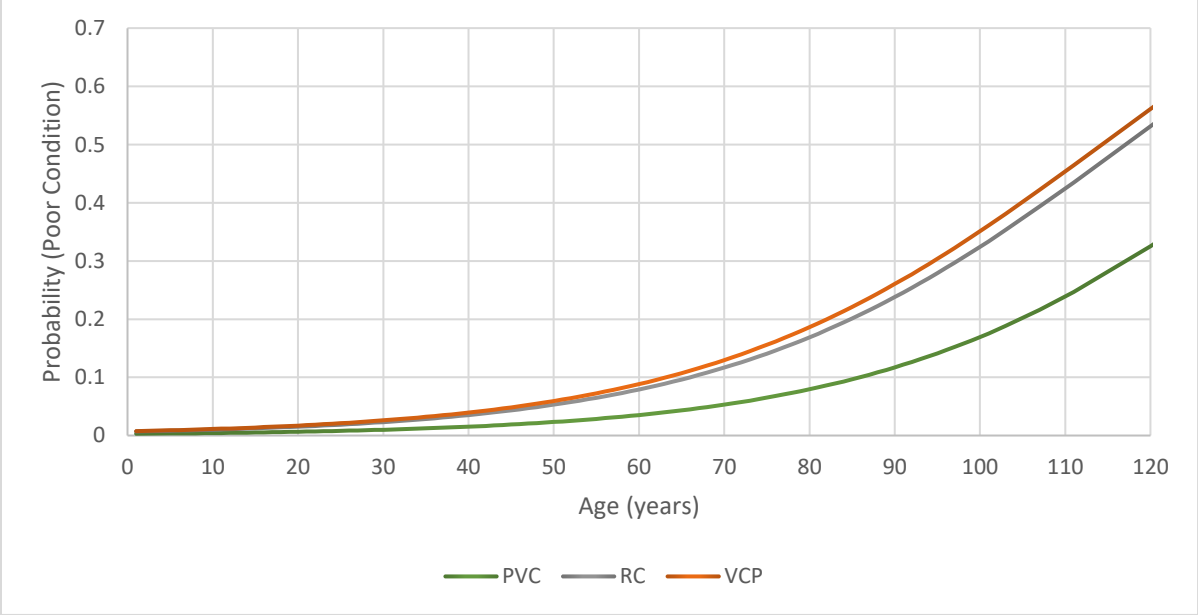

Figure 5-6 Deterioration Curve for Sewer Pipes with Different Materials Buried in Sand

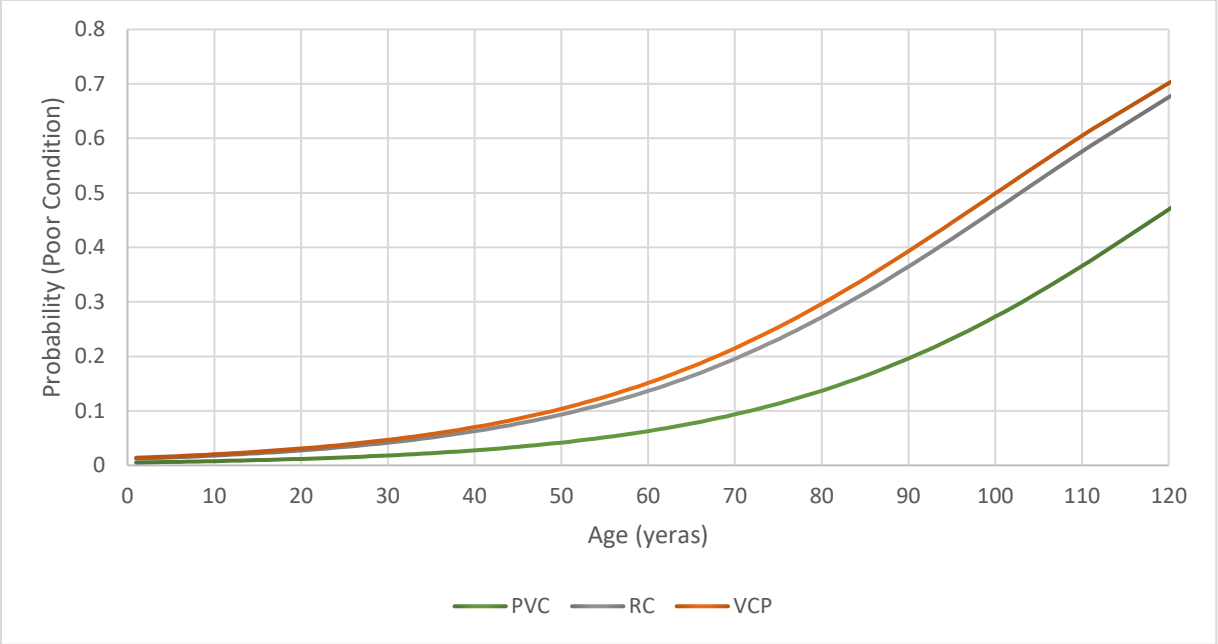

Figure 5-7 Deterioration Curve for Sewer Pipes with Different Materials Buried in Clay

As may be expected, the likelihood of sewer pipes being in poor condition increased as they grew older. Every pipe material has a specific functional life, and its physical characteristics alter as it ages. Therefore, it follows that sewer pipes' corrosion rate increases with age. The previous finding supports the findings of various studies described in section 2.4.2.

It can be seen that in both soils, the deterioration rate of PVC pipes is the lowest and for VCP pipes is the highest. Varying materials used to construct sewer pipes react differently to environmental conditions, such as soil type. For instance, clay pipes can withstand acids well, whereas concrete pipes resist abrasion. PVC and other plastic pipes are resistant to acidic and alkaline wastes, but heavy loads might cause them to distort excessively (Malek Mohammadi, 2019; Singh and Adachi, 2013). The above results support the findings of the investigations reported in section 2.4.3.

Figures 5-8 and 5-9 show the effect of length on the degradation of sewer pipes. Figure 5-8 illustrates the effect of length on a 48-year-old pipe constructed with different materials buried in the sand. On the other hand, Figure 5-9 depicts the effect of length on the same pipe constructed from VCP (which is the maximum in the available dataset) buried in different soil types.

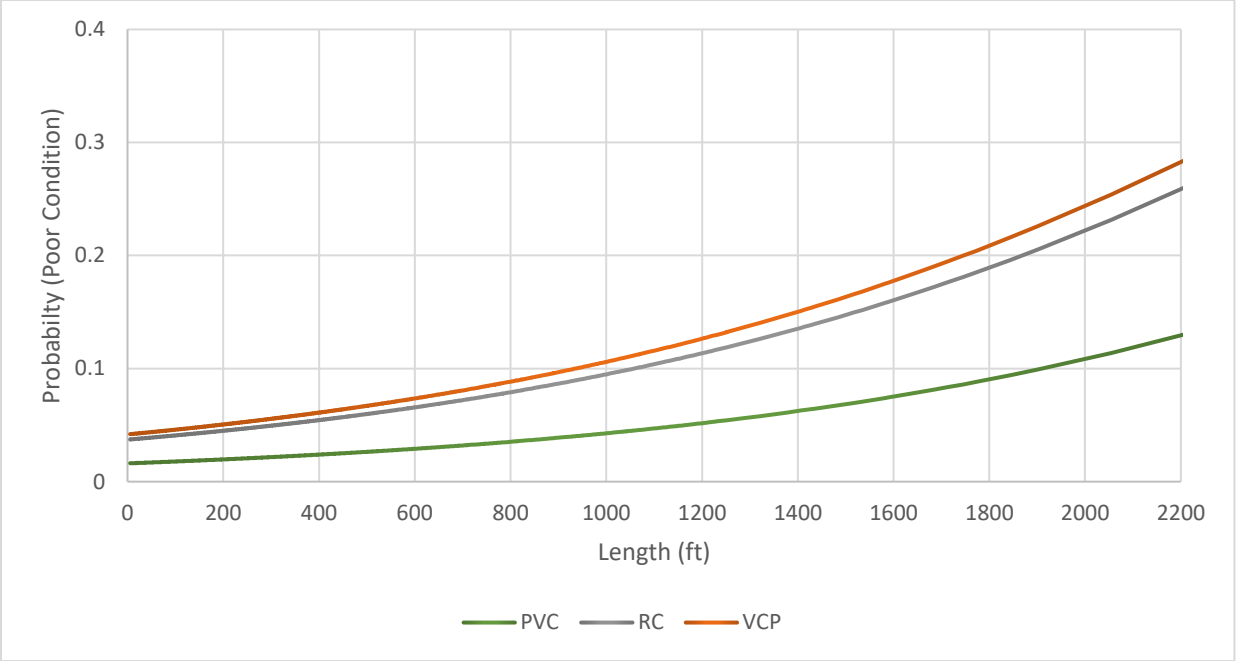

Figure 5-8 Effect of Pipe Length on Condition of a 48-year-old Pipe Made by Different Materials

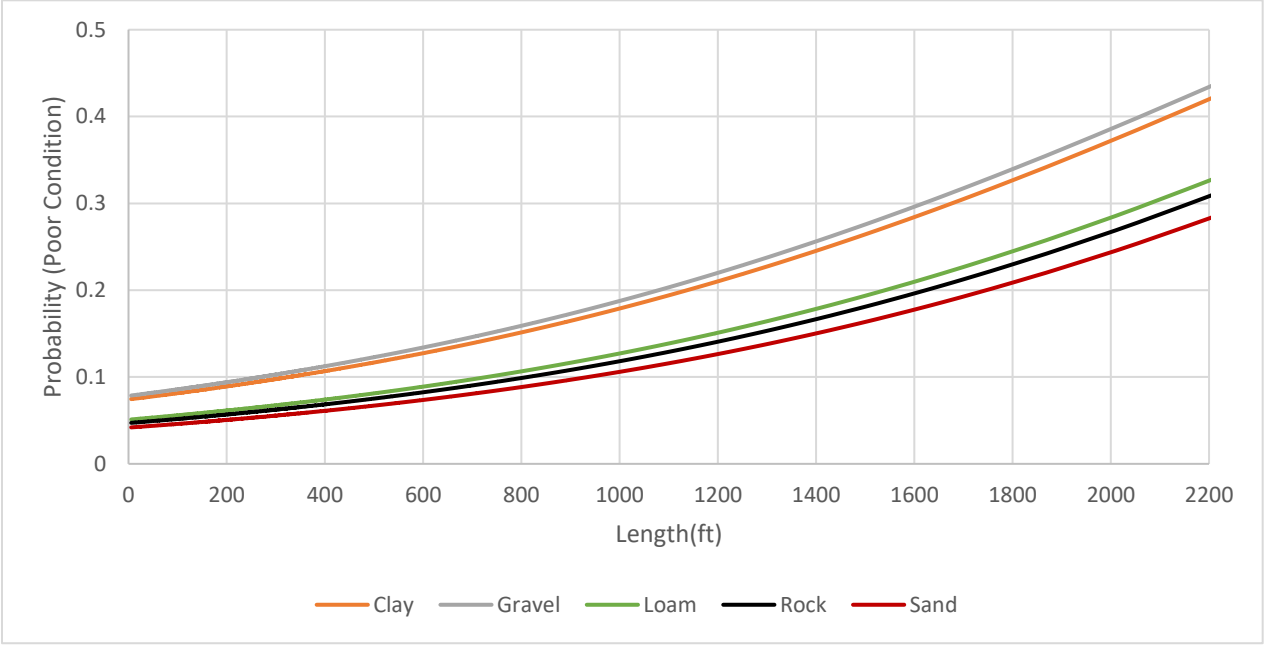

Figure 5-9 Effect of Pipe Length on Condition of a 48-year-old Pipe Buried in Different Soil Types

As is evident in the Figures above, the probability of being in poor condition increases with the pipes' length. The primary infiltration point is at pipe joints, and longer pipes have more potential failure points and areas, particularly at joints. Because defects are more likely to occur in longer sewer pipes, they often deteriorate faster. Additionally, blockages and sediment accumulation are more likely to occur in longer lines, speeding up sewer pipes' deterioration. This finding is consistent with results of some studies mentioned in section 2.4.5.

Figures 5-10 to 5-12 show the effect of different types of soils on a 283-ft sewer pipe constructed with various materials. As explained in chapter 4, there are 5 types of soil in the available dataset, including sand, clay, loam, gravel, and rock.

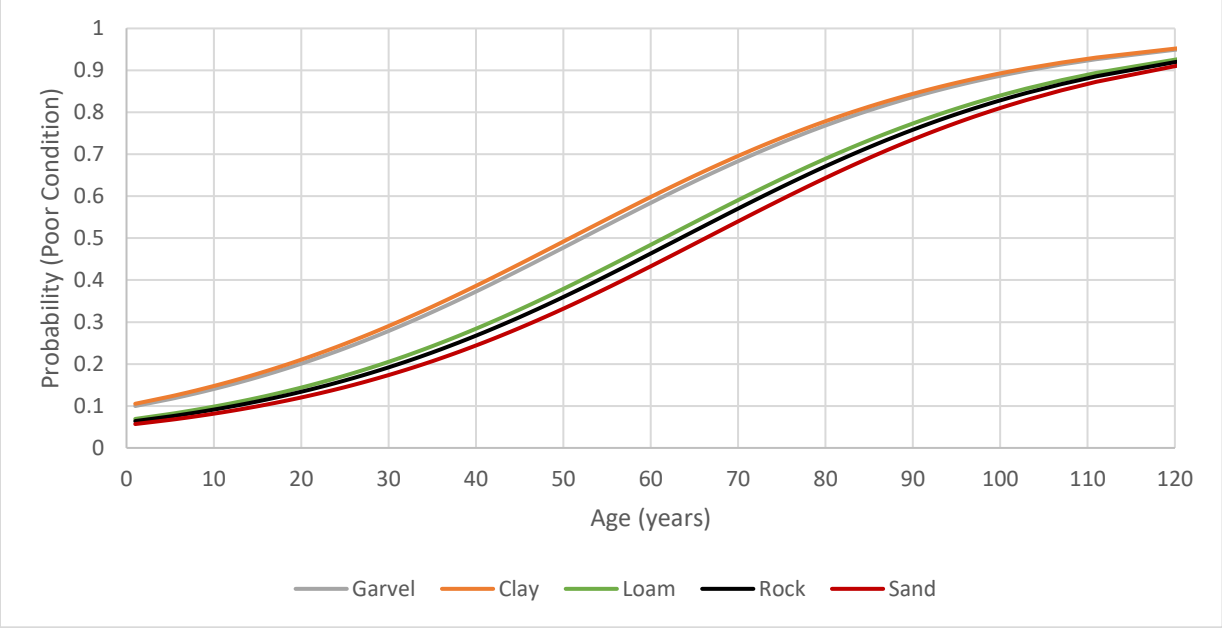

Figure 5-10 Effect of Surrounding Soil Type on Condition of a 283-ft Pipe made by PVC

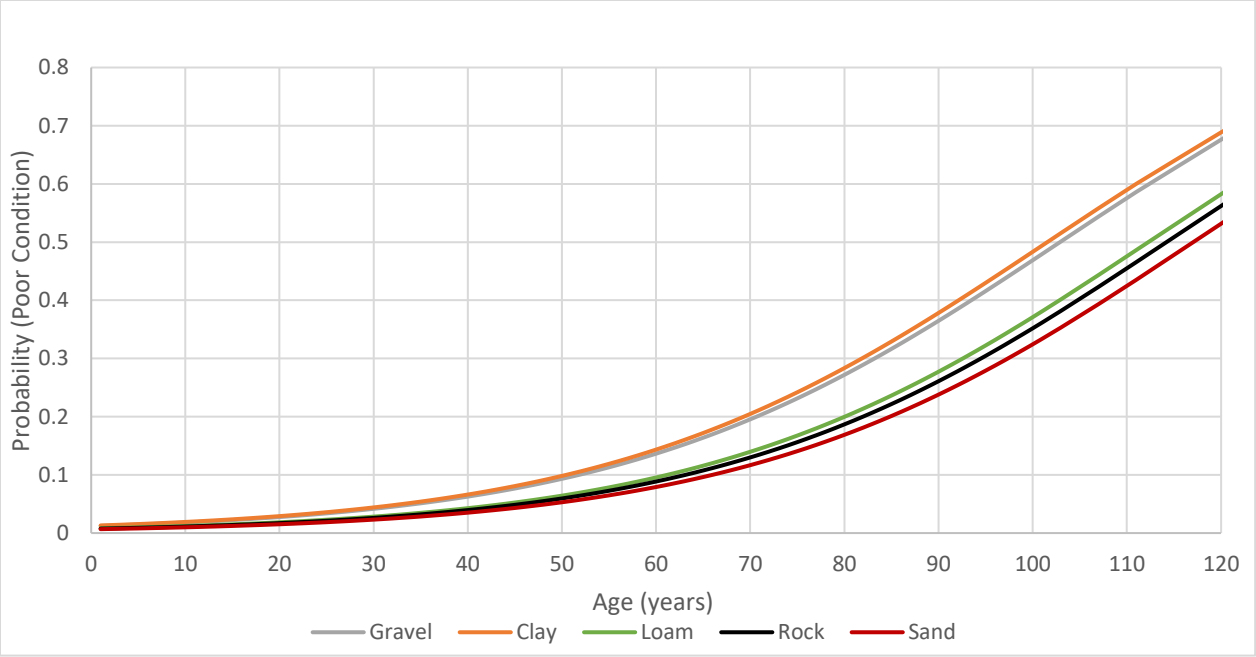

Figure 5-11 Effect of Surrounding Soil Type on Condition of a 283-ft Pipe made by RC

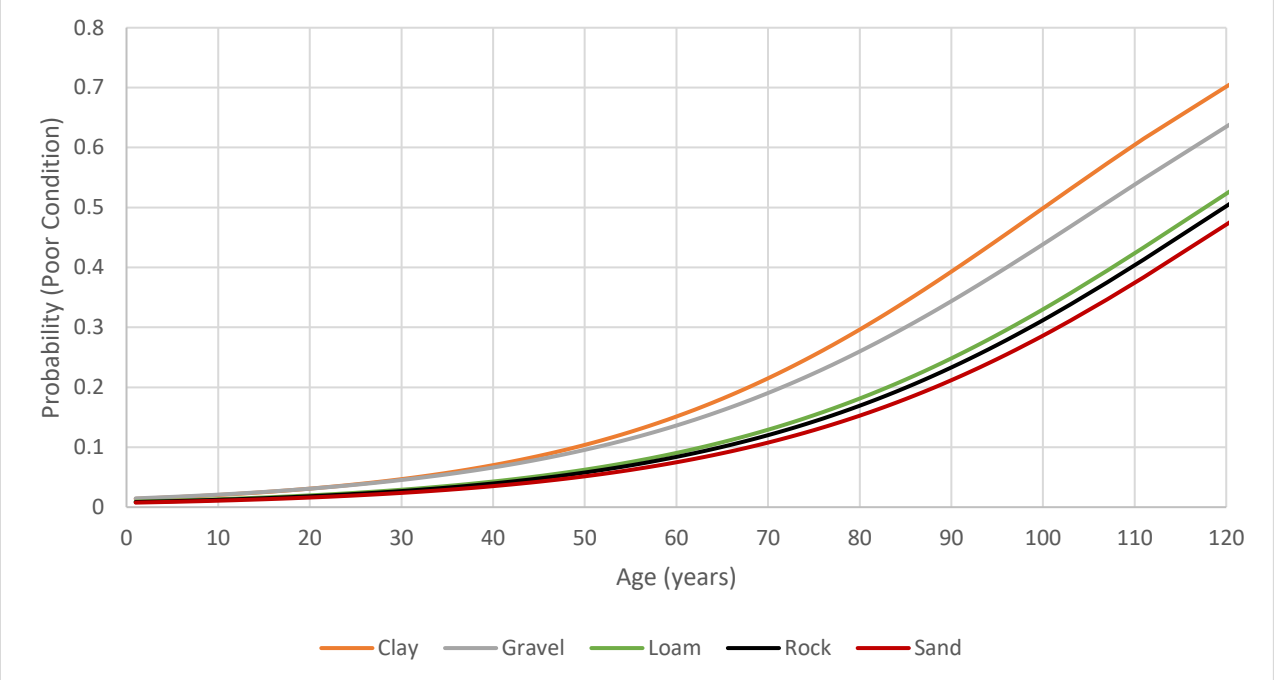

Figure 5-12 Effect of Surrounding Soil Type on Condition of a 283-ft Pipe made by VCP

In all figures above, the pipes surrounded by sand have the lowest deterioration rate than pipes surrounded by other soil types; on the other hand, pipes buried in clay have the highest deterioration rate. The results of investigations by Davies et al. (2001) and O'Reilly et al. (1989) support this finding. Graphs show that sewers surrounded by clayey soil have more problem. It could be for two reasons.

Firstly, clayey soils have typically more moisture content (high plasticity index) which is an indicator of high potential of absorbing water. Consequently, they are classified as frost-susceptible soils. According to Najafi and Gokhale (2022), “Frost heave is defined as the vertical expansion of soil caused by soil freezing and ice lens formation. Differential heave causes sections of pipe to experience non-uniform displacements and this results in forceful flexural stresses. Also, uniform heaving may cause a problem where pipe joints are not subjected to movement. In this case the pipe will experience bending stresses. So, failure of pipe joints could be the result of heave process” (p. 53).

Secondly, clay generally has lower internal fraction rather than sand and gravel and consequently lower shear strength. This characteristic results in lower soil-pipe interaction in clayey soils. Because the shear strength of the interaction might change the pipeline's degree of mobility and hence increase its displacement, soil-pipe interaction is crucial. High soil-pipe interaction prevents the pipe from contracting, which causes the axial stress to increase.

5.5.2 Multinomial Logistic Regression

The first method developed by Python using resampled dataset was Multinomial Logistic Regression. As discussed earlier, Multinomial logistic regression can be used as an extension of binary logistic regression when the dependent variable is categorical and contains more than two levels.

5.5.2.1 Training the Model

Firstly, the dependent and independent variables were defined. Then, using the train_test_split method of Python, 80% of the data were set for training the model and 20% for testing. The sklearn.linear_model.LogisticRegression library of Python was used to implement the Logistic Regression model.

5.5.2.2 Results and Discussions

To be consistent with other AI methods implemented in this study, the evaluation metric for the developed Multinomial Logistic Regression model was selected to be the confusion matrix instead of the

classification table. Figure 5-13 shows the confusion matrix of this model. High values on non-diagonal cells of the confusion matrix below demonstrate that misclassification was high in the developed MLR prediction model. Specifically, pipes with condition ratings of 3 had significant misclassification with pipes having condition ratings of 2 and 4.

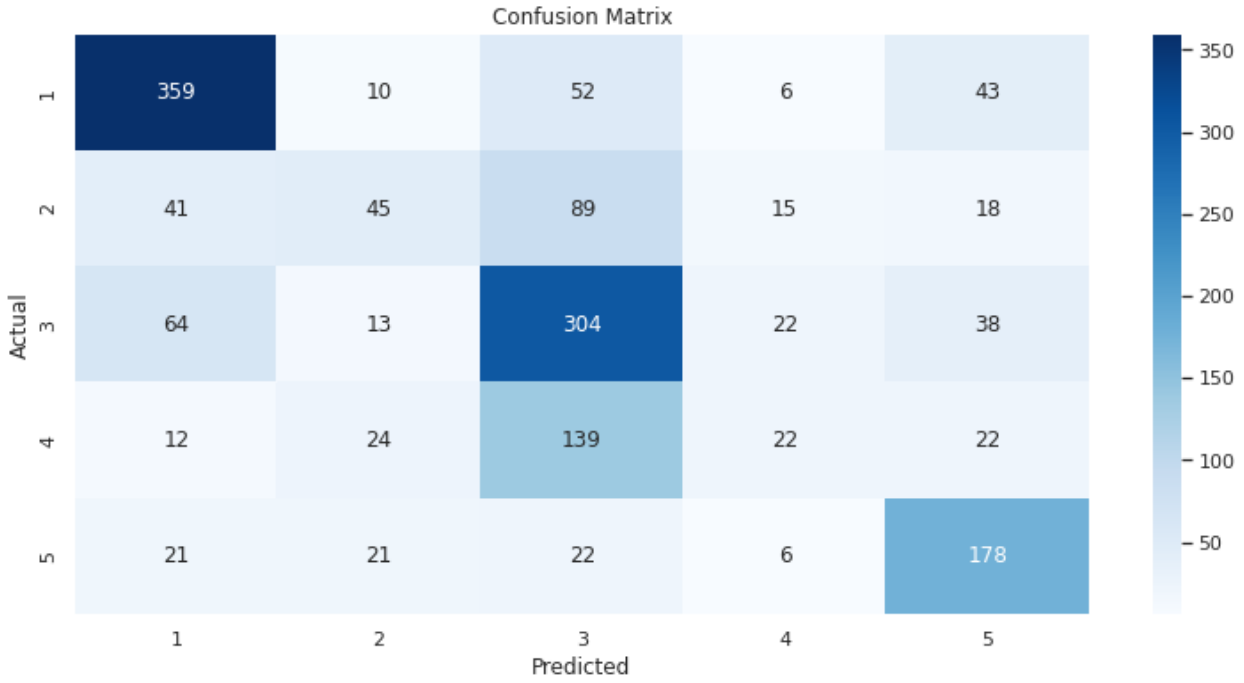

Figure 5-13 Confusion Matrix of Multinomial Logistic Regression Model

As described in chapter 3, another important metric to evaluate the performance of the developed modes is the ROC curve, including the AUC criteria. As it can be seen in Figure 5-14, the ROC curve of classes 2,3, and 4 are not close to the upper left corner. However, detailed measurements should be investigated to assess the model's effectiveness. They are calculated based on the confusion matrix and are shown in Table 5-9.

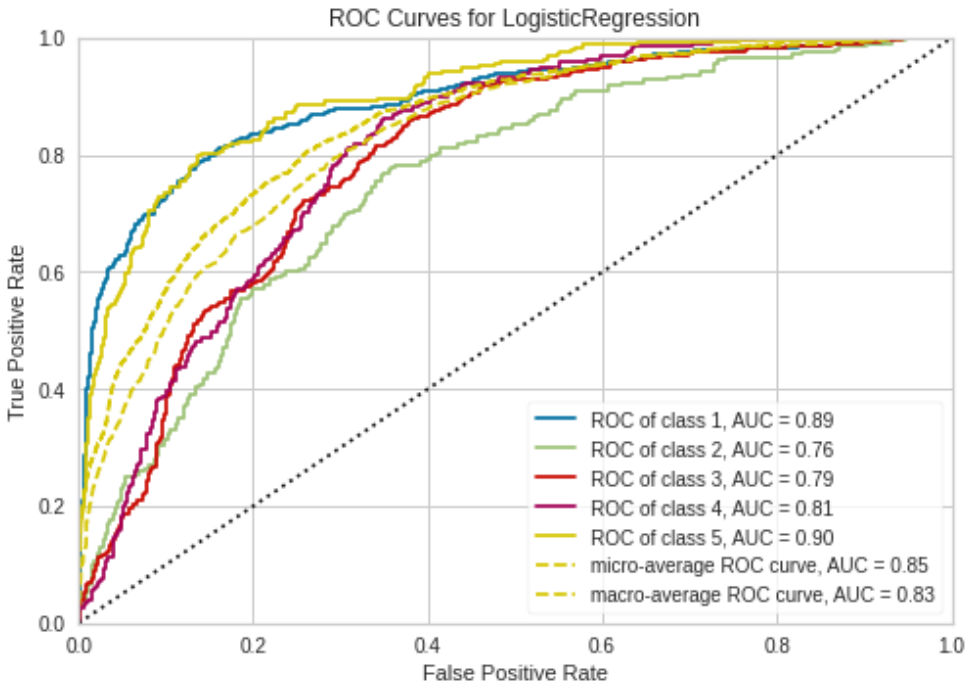

Figure 5-14 ROC Curves for Multinomial Logistic Regression Model

Table 5-9 Precision, Recall, and F1-Score Metrics for Multinomial Logistic Regression Model

Condition Rating	Precision	Recall	F1 Score
1	0.72	0.76	0.73
2	0.39	0.23	0.29
3	0.50	0.68	0.57
4	0.30	0.10	0.15
5	0.59	0.71	0.64
Macro - Average	*0.49*	*0.48*	*0.48*

F1-scores of pipes with condition ratings of 2 and 4 are very low, indicating the weak performance of the developed model in predicting these classes. A precision value of 0.59 for class 5, which is an

important class in sewer prediction modeling, shows that the model only correctly identified 59% of pipes with a condition rating of 5 out of all the pipes classified as pipes with a condition rating of 5. Finally, the macro-average F1-score of 0.48, the summary result of the model's efficiency, states that the developed MLR model was unsuccessful.

5.6 KNN

5.6.1 Training the Model

As explained before, 5-fold cross validation method was used to develop the models including KNN. It randomly selected 80% of data for train and 20% for testing the model. The Python Scikit-learn library's K neighbors classifier parameters are shown in Table 5-10.

Table 5-10 Parameters of the Developed KNN Model

Parameters	Description
n_neighbors	Specify number of neighbors: 7
weights	weight function used in prediction: uniform, distance
algorithm	Algorithm used to compute the nearest neighbors: auto, ball_tree
leaf_size	This parameter is estimated by ball_tree

To maintain the model's consistency, the weight parameter was set uniformly. Some parameters, including leaf size, were set to default values to create the KNN model. The nearest neighbors were calculated using the auto algorithm because this function tries to discover the best algorithm. The most crucial step in the KNN model's development is determining the number of neighbors (K). When smaller values for are chosen, there is generally a high danger of overfitting. This parameter can be determined manually. So, the model was run using different K values from 3 to 11. The K value should be odd because of the voting issue during the KNN model development. Therefore, using numbers 3, 5, 7, 9, and 11 as the value of K, various KNN models were implemented, and the highest accuracy was achieved when the number of neighbors was 7. Therefore, the model was developed with 7 neighbors, and the testing results are presented in the next section.

5.6.2 Results and Discussions

The performance of the k-NN model developed using resampled data is reviewed in this section. The number of instances in each class in the oversampled dataset matches the dominant class. As seen in Figure 5-15, a confusion matrix was produced using the generated k-NN model. It should be noted that the confusion matrix below is based on the testing part of the developed model and is the validation of the model.

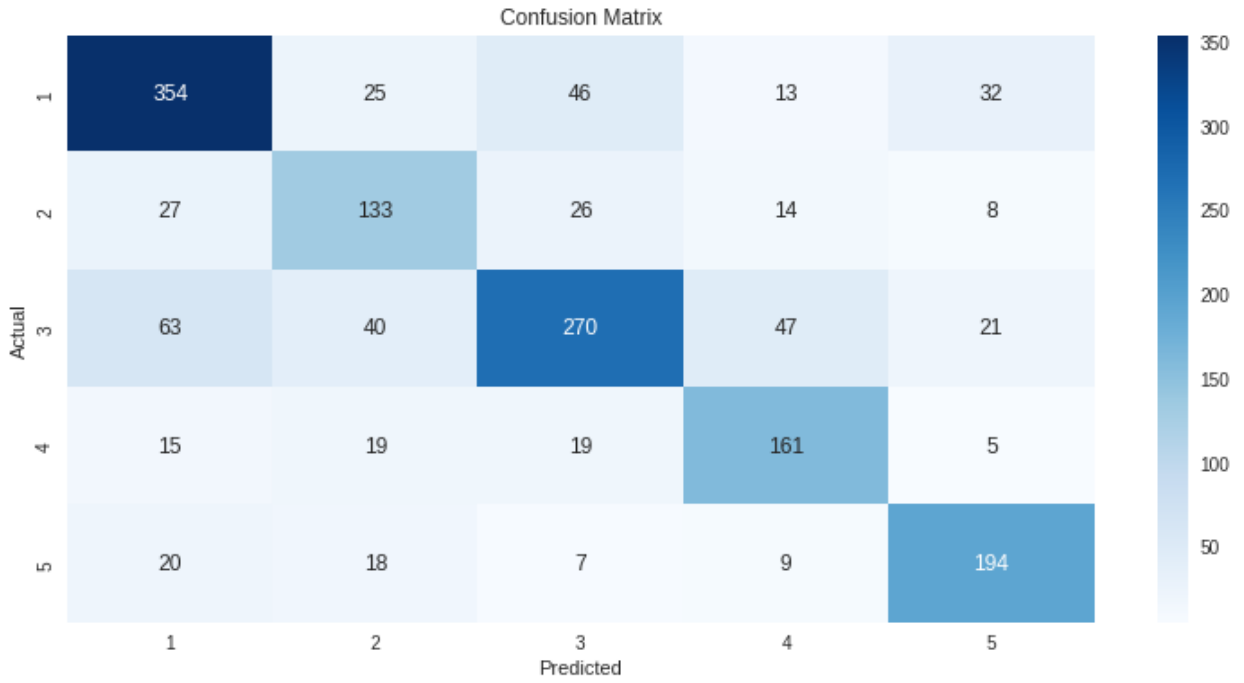

Figure 5-15 Confusion Matrix of KNN Model

As described in section 5.4.3, the sum of the values in the matrix should equal 1588, which is valid for the above table. Furthermore, as explained in chapter 3, the confusion matrix's main element is diagonal cells showing the model's performance for each class. Higher numbers and darker backgrounds in these cells show the better performance of the model for each class. Consequently, lower numbers and brighter backgrounds in other cells are preferred since they are misclassified instances.

Here, to clarify the subject, an example of the status of condition rating of 1 and condition rating of 2 from the above matrix is explained. The value of 354 on the top left of the matrix illustrates the number of sewer pipes that actually have the condition rating of 1 and have been predicted correctly. The value of 25 on the second cell of the first row shows the number of pipes that actually are in the condition rating of 1,

but the model has predicted them wrongly as pipes with a condition rating of 2. On the other hand, the value of 27 on the second cell of the first column depicts the number of pipes that actually are in the condition rating of 2, but the model has predicted them wrongly as pipes with a condition rating of 1.

ROC curves were plotted as shown in the Figure 5-16 in order to better understand the performance of the k-NN model. As discussed in section 3.5.2, the model has higher overall accuracy when the ROC curve is closer to the upper left corner, and consequently, the AUC value is close to 1.

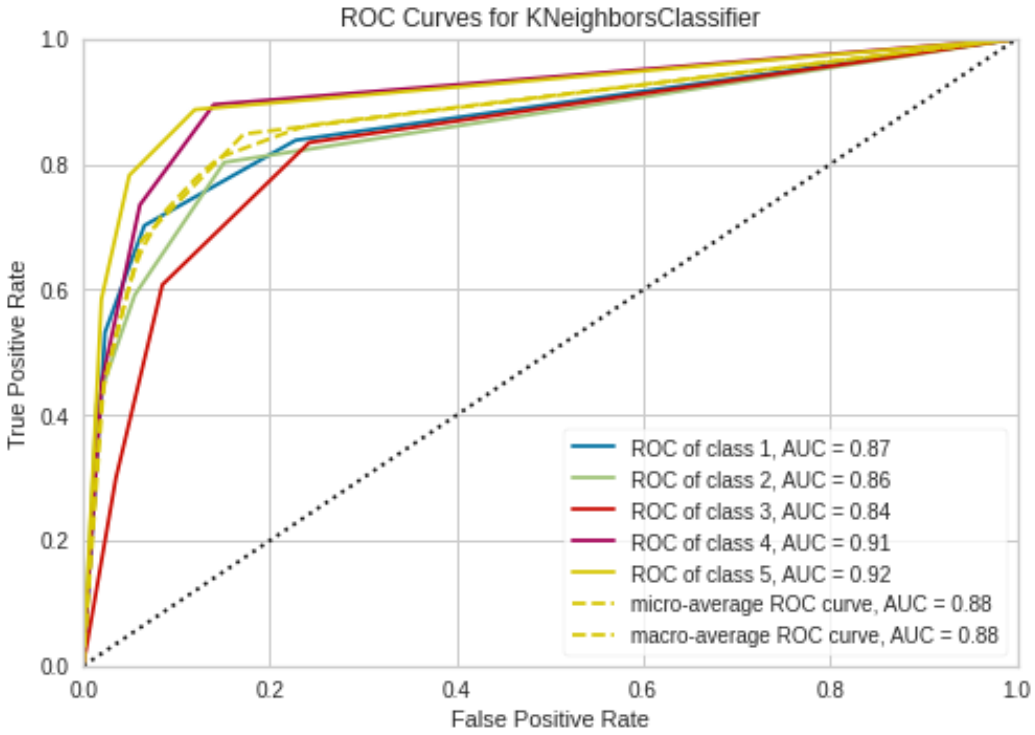

Figure 5-16 ROC Curves for k-NN Model

AUC values were closer to 1 for curves corresponding to condition ratings of 4 and 5, demonstrating the model's high accuracy in predicting these classes. AUC was found to be 0.87, 0.86, and 0.84 for classes 1, 2, and 3, respectively.

Evaluation metrics, including precision, recall, and finally, F1-score, were calculated using generated confusion matrix and based on explanations of section 3.5.6. Table 5-11 displays the recall, precision, and F1-score of the developed KNN model.

Table 5-11 Precision, Recall, and F1-Score Metrics for k-NN Model

Condition Rating	Precision	Recall	F1 Score
1	0.74	0.75	0.74
2	0.57	0.64	0.60
3	0.74	0.61	0.67
4	0.66	0.74	0.70
5	0.75	0.78	0.76
Macro - Average	*0.70*	*0.71*	*0.70*

To explain the table above in simple word, the metrics for pipes with a condition rating of 4 is explained. A precision value of 0.66 shows that the model correctly identified 66% of pipes with a condition rating of 4 out of all the pipes classified as pipes with a condition rating of 4. The recall value of 0.74 shows that the model correctly identified 74% of pipes having a condition rating of 4 out of all the pipes actually having a condition rating of 4. Finally, the F1 score, by combining them into a single measure, evaluates the model's effectiveness. An F1 score of 0.70 shows the ability of the model to predict pipes in a condition rating of 4.

F1 closer to 1 shows better performance of the model. An overall macro-average F1-score of 0.70 shows an acceptable performance of the developed KNN model for the available sewer pipes dataset. It was found that the KNN model had better performance for pipes with a condition rating of 5 (F1= 0.76) rather than others and had the lowest accuracy in predicting pipes with a condition rating of 2 (F1=0.60).

5.7 Tree-Based Models

5.7.1 Decision Tree

To compare the performance of different machine learning methods to achieve the most accurate model, one of this study's main goals, the tree-based models, was developed. In the beginning, the regular Decision Tree classifier was tried. The concept of this method is described in section 3.4.2.

5.7.1.1 Training the Model

The main parameters in training a Decision Tree model are maximum depth and splitting criterion. For the criterion, the Gini index was selected as default. For maximum depth, researchers suggest a trial-and-error process or depth equal to the number of dataset's attributes which means the number of variables (Geron, 2017; Müller and Guido, 2016). A trial-and-error process was done, and the maximum depth of 11 was selected as the best.

5.7.1.2 Results and Discussions

This part evaluates the effectiveness of the created Decision Tree model. The resultant Decision Tree model was used to create a confusion matrix, as shown in Figure 5-17. The interesting point is that despite the model's good performance regarding classes 1 and 3, the misclassification between these two is a little high.

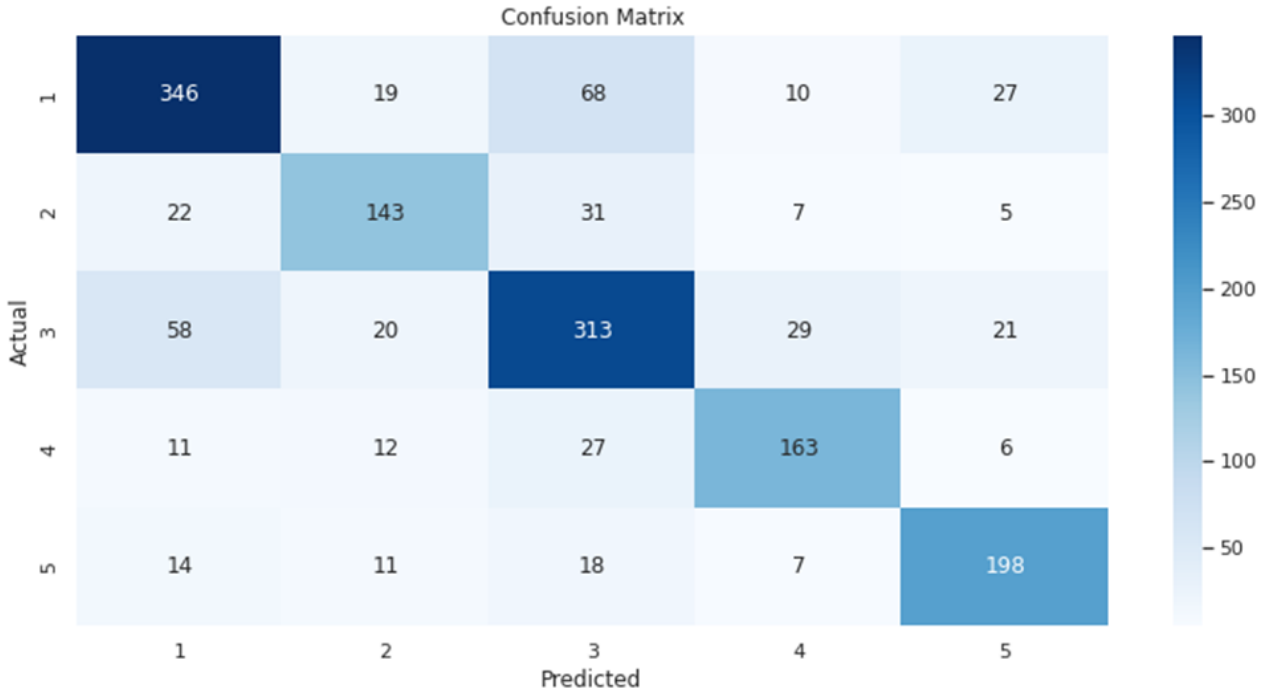

Figure 5-17 Confusion Matrix of Decision Tree Model

Figure 5-18 illustrates the ROC curve of the model. According to the figure, the model has more AUC for pipes with poor condition ratings (PACP 4 and 5) than for other classes, which is a good sign. However, other metrics such as precision, recall, and F1 scores must be estimated to evaluate the performance of a prediction model in detail.

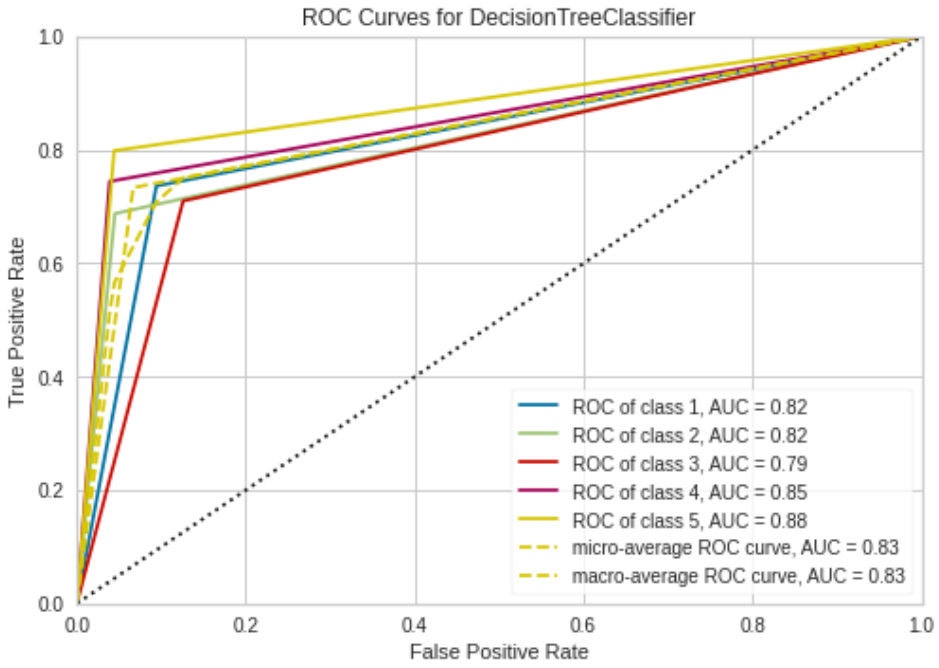

Figure 5-18 ROC Curves for Decision Tree Model

According to Table 5-12, the developed model performs less well for pipes with a condition rating of 2 than it does for other pipes. It might be because there are only a few pipes in the dataset in this situation. Additionally, it can be observed that pipes in class 5 have the greatest F1-scores, and the model can accurately forecast them. Overall, the Decision Tree classifier's model, which has an F1-score of 0.73, outperforms the KNN model.

Table 5-12 Precision, Recall, and F1-Score Metrics for Decision Tree Model

Condition Rating	Precision	Recall	F1 Score
1	0.74	0.74	0.74
2	0.70	0.69	0.69
3	0.69	0.71	0.70
4	0.75	0.74	0.74
5	0.77	0.80	0.78
Macro - Average	*0.73*	*0.74*	*0.73*

5.7.2 Random Forest

The first ensemble of the Decision Tree algorithm to be tested was the Random Forest. As discussed earlier, the Random Forest algorithm is based on the Bagging approach. It constructs different decision trees and reaches to best result by taking the average of the final results of the trees. The randomness in this method means that each tree is constructed with a random dataset and a random variable.

5.7.2.1 Training the Model

The main parameters for this method which should be assigned are number of decision trees (n_estimators) and the number of features to be analyzed in each tree (n_features). A higher number of constructed trees usually leads to higher accuracy (Geron, 2017). Therefore, after the trial-and-error process, the number of trees was set to 100. So, 100 distinct trees were constructed using 100 different datasets randomly selected from the original dataset.

Next, in each tree the algorithm randomly selects a subset of the features (independent variables), and it looks for the best possible test involving one of these features. The number of features that are selected is controlled by the max_features parameter. Max_feature is a critical parameter in this method. If we set it equal to the number of independent variables (which in our case is 9), that means that each tree can look at all variables in the dataset, and no randomness will be injected into the feature selection (though the randomness due to the bootstrapping of data remains). If we set it to 1, the trees have no choice on which feature to test and can only search over for the randomly selected variable.

As a result, a high max_features value indicates that the trees in the random forest will be relatively similar and easily fit the data using the most distinguishing features. The trees in the random forest will be significantly different if the max_features are low, and each tree may need to be very deep to adequately fit the data (Müller and Guido, 2016). It is possible to leave the selection of the best value for this parameter to the algorithm itself. Consequently, this parameter was decided to be determined by the algorithm, and it had been a number between 1 and 9.

5.7.2.2 Results and Discussions

Finally, the prediction model for sewer pipes condition assessment by RF algorithm was trained via Python, and the results are shown in this section. Figure 5-19 depicts the confusion matrix of the testing

part of the developed model. Relatively low values in the non-diagonal cells are a sign of the model's outstanding performance.

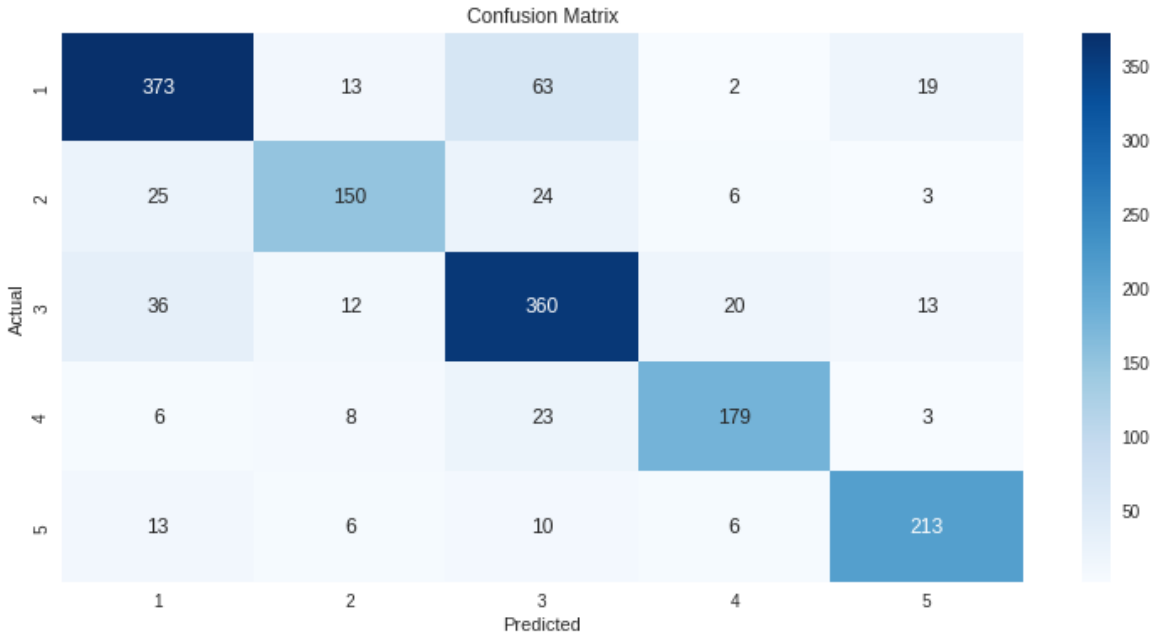

Figure 5-19 Confusion Matrix of Random Forest Model

ROC curves for all classes were plotted as shown in Figure 5-20. It can be seen that the AUC for ROC curves of condition ratings 4 and 5 are close to the unit. Generally, in sewer pipe prediction modeling, high values in metrics related to minority conditions (PACP 4 and 5) show the high efficiency of the model. ROC curves of other classes are also in excellent condition.

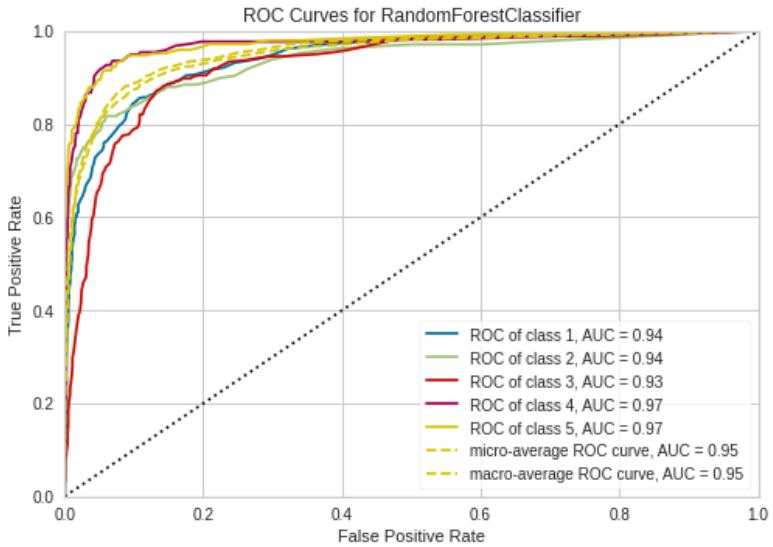

Figure 5-20 ROC Curves for Random Forest Model

Table 5-13 illustrates other metrics for the developed RF model. F1-scores of classes 1, 4, and 5 are high, indicating excellent performance in predicting them. The precision of 0.82 for class 1 reveals that the model correctly identified 82% of pipes with a condition rating of 1 out of all the pipes classified as pipes with a condition rating of 1. Since a high quantity of pipes belongs to this class, this precision value is considered high accuracy. Furthermore, a recall value of 0.86 for class 5 indicates that the model correctly identified 86% of pipes having a condition rating of 5 out of all the pipes actually having a condition rating of 5. It is a symbol of low misclassification of the developed model. Finally, the overall macro-average F1-score of 0.8 shows that the model had very good performance in predicting all classes and outperforms the KNN and Decision Tree models.

Table 5-13 Precision, Recall, and F1-Score Metrics for Random Forest Model

Condition Rating	Precision	Recall	F1 Score
1	0.82	0.80	0.81
2	0.80	0.72	0.76
3	0.75	0.81	0.78
4	0.84	0.82	0.83
5	0.85	0.86	0.85
Macro - Average	*0.81*	*0.80*	*0.80*

5.7.3 AdaBoost

AdaBoost is the first algorithm among the Boosting algorithms to be tried. To compare their findings to those of other tree-based models, boosting techniques were applied. AdaBoost simply calculates the predictions from each predictor and weighs them using the predictor weights (the higher the weight of the predictor, the more accurate the predictor is). The class that receives the majority of weighted votes is the predicted class (Geron, 2017). The number of trees and the learning rate, which controls how much each tree is permitted to correct the errors of the previous trees, are the two key parameters of the boosted models. These two factors are related because more trees are required to construct a model with a same

level of complexity at a lower learning rate. In boosting models, adding more estimators makes the model more complex, which might result in overfitting, in contrast to random forests where having more estimators (predictors)(trees) is always beneficial. It is usual practice to fit several estimators based on the time and memory budget before looking through various learning rates. A common practice is to fit number of estimators depending on the time and memory budget, and then search over different learning rates (Müller and Guido, 2016). By these explanations, the number of predictors was set to 50, and the learning rate was set to 1 as default. It should be noted that decreasing the learning rate would cause an increase in the number of trees.

The confusion matrix of the developed model is shown in Figure 5-21. The values in non-diagonal cells are high, illustrating relatively high misclassification among different classes. The misclassification between classes 3 and 4 is higher than in others.

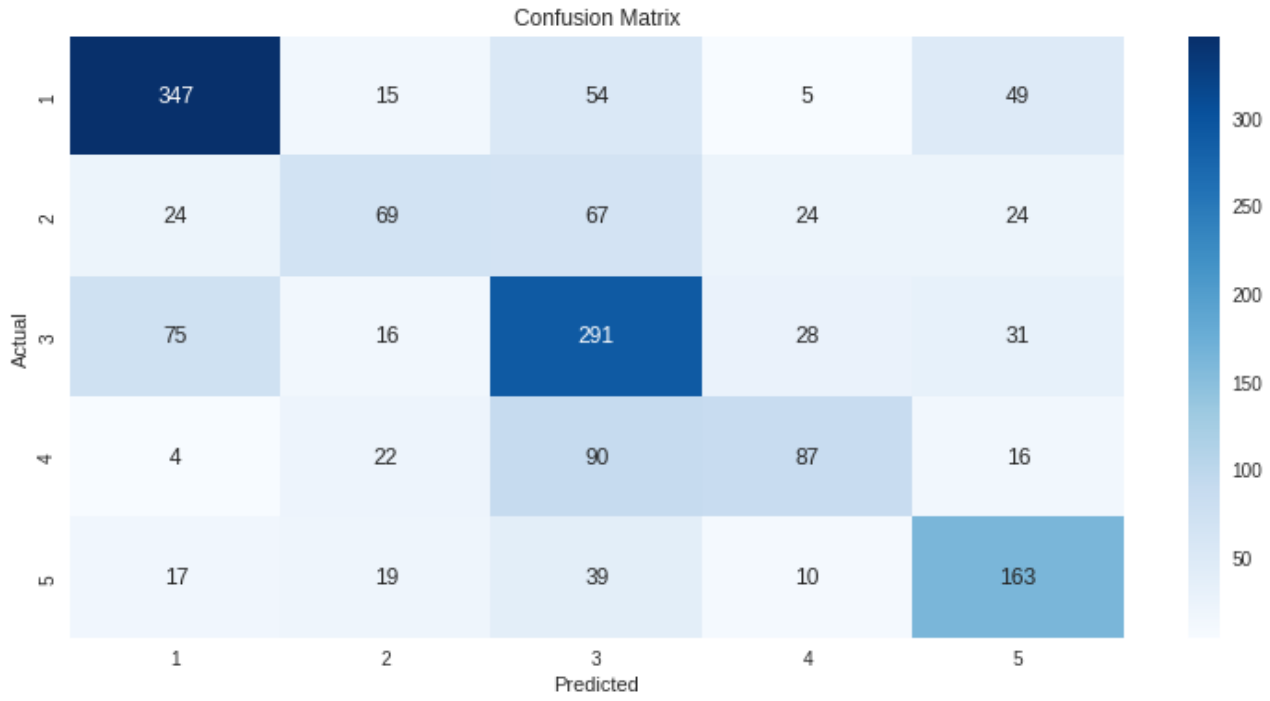

Figure 5-21 Confusion Matrix of AdaBoost Model

The ROC curve of this model is depicted in Figure 5-22. It can be seen that the AUC of classes 2 and 3 are relatively low. Other metrics should be discussed to evaluate the performance of the model.

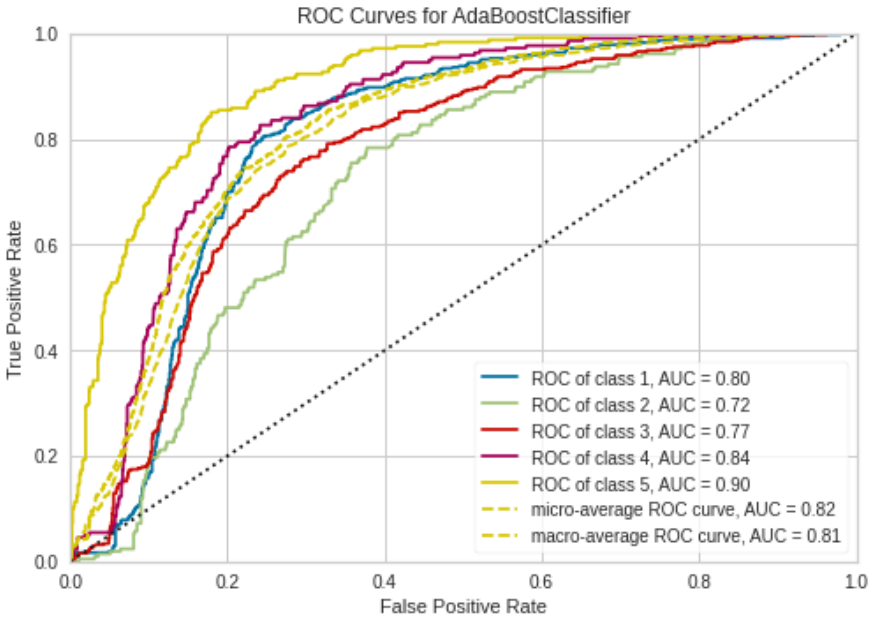

Figure 5-22 ROC Curves for AdaBoost Model

Precision, recall, and F1-score of all classes are shown in Table 5-14. Low F1-score for pipes with a condition rating of 2 and 4 show the weak performance of the developed model for these classes. Also, the precision and recall values of the model for pipes with a condition rating of 5, which has always been a sign of the model's performance, are low. Finally, the overall macro-average F1-score of 0.57 indicates that AdaBoost was not a suitable model for the available dataset.

Table 5-14 Precision, Recall, and F1-Score Metrics for AdaBoost Model

Condition Rating	Precision	Recall	F1 Score
1	0.74	0.74	0.74
2	0.49	0.33	0.39
3	0.54	0.66	0.60
4	0.56	0.40	0.47
5	0.58	0.66	0.62
Macro - Average	*0.58*	*0.56*	*0.57*

One significant drawback to this sequential learning technique is that it cannot be parallelized (or only partially) since each predictor can only be trained after the previous predictor has been trained and evaluated. As a result, it does not scale as well as bagging. The other disadvantage of boosting is that it is sensitive to outliers since every classifier must fix the errors in the predecessors. Thus, the method is too dependent on outliers (Geron, 2017; Müller and Guido, 2016). These two disadvantages could be the reason for the weak performance of the developed AdaBoost model.

5.7.4 Gradient Boosting Tree

Another Boosting approach tried in this study was Gradient Boosting Tree. As described in section 3.4.4.2, this method operates similarly to AdaBoost. The only difference is that this method modifies the residual errors instead of modifying instance weights. Again, the main parameters are the number of trees and the learning rate. The learning rate was set at one as default. According to Geron (2017), “in order to find the optimal number of trees, you can use early stopping. simple way to implement this is to use the staged_predict code in the Python: it returns an iterator over the predictions made by the ensemble at each stage of training (with one tree, two trees, etc.). The following code trains a GBT ensemble with 120 trees, then measures the validation error at each stage of training to find the optimal number of trees, and finally trains another GBT ensemble using the optimal number of trees” (p. 208).

The confusion matrix of the developed Gradient Boosting Trees model is shown in Figure 5-23. The values on diagonal elements are evidence of a relatively proper result for the developed model, specifically for pipes with condition ratings of 1 and 3.

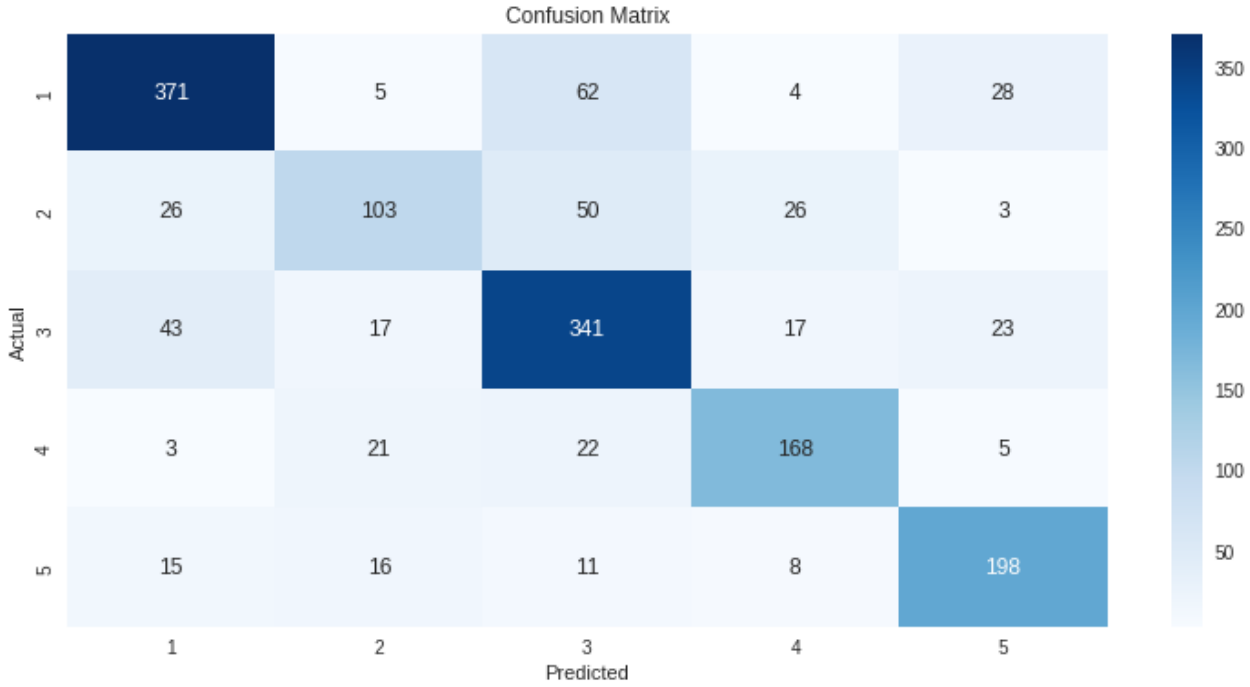

Figure 5-23 Confusion Matrix of GBT Model

Another tool to evaluate the developed GBT model was the ROC curve. It is shown in Figure 5-24. Unlike the AdaBoost model, AUC values for all classes are higher than 0.9, which indicates that the model had a better performance than the previous boosting model.

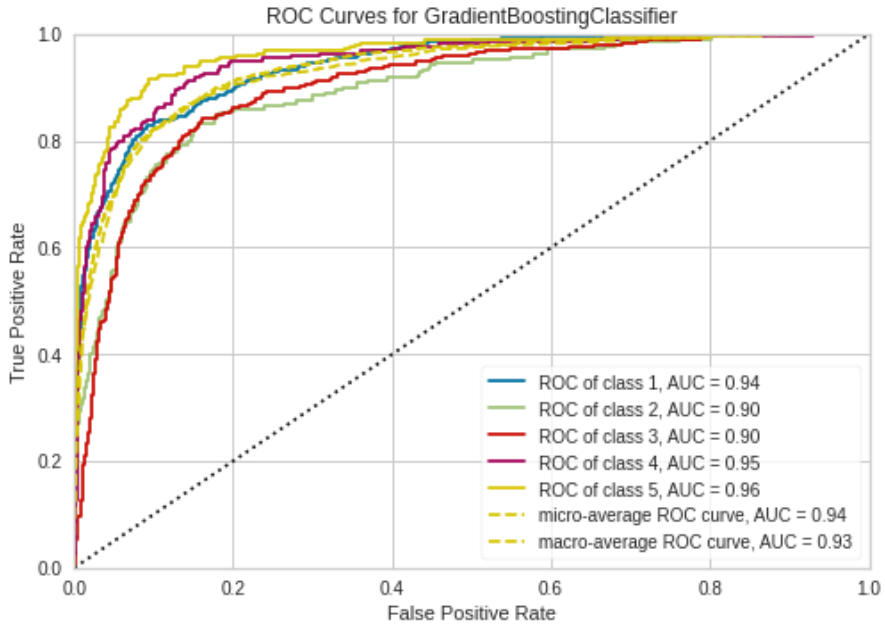

Figure 5-24 ROC Curves for GBT Model

Table 5-15 explains other evaluation metrics for the GBT model. Except for pipes with a condition rating of 2, precision and recall values of other classes are acceptable. The calculated F1-score shows that this model is much more accurate than the AdaBoost model.

Table 5-15 Precision, Recall, and F1-Score Metrics for GBT Model

Condition Rating	Precision	Recall	F1 Score
1	0.81	0.79	0.80
2	0.64	0.50	0.56
3	0.70	0.77	0.73
4	0.75	0.77	0.76
5	0.77	0.80	0.78
Macro - Average	*0.73*	*0.73*	*0.73*

One reason for the better performance of GBT than AdaBoost could be the optimal number selection of trees which was set to be automatically done. The other reason could be that GBT uses a loss function (such as squared-error) to correct the errors of the prior tree, which had been more effective than the AdaBoost approach, which uses modifying instances weights to correct the previous trees.

5.7.5 XGBoost

As mentioned in section 3.4.4.3, the XGBoost method is the same as GBT but faster. This model's confusion matrix is shown in Figure 5-25. For pipes with condition ratings of 1 and 3, which are the most numerous, the non-diagonal cell values show no significant misclassification. This sign demonstrates that the model is operating within acceptable bounds.

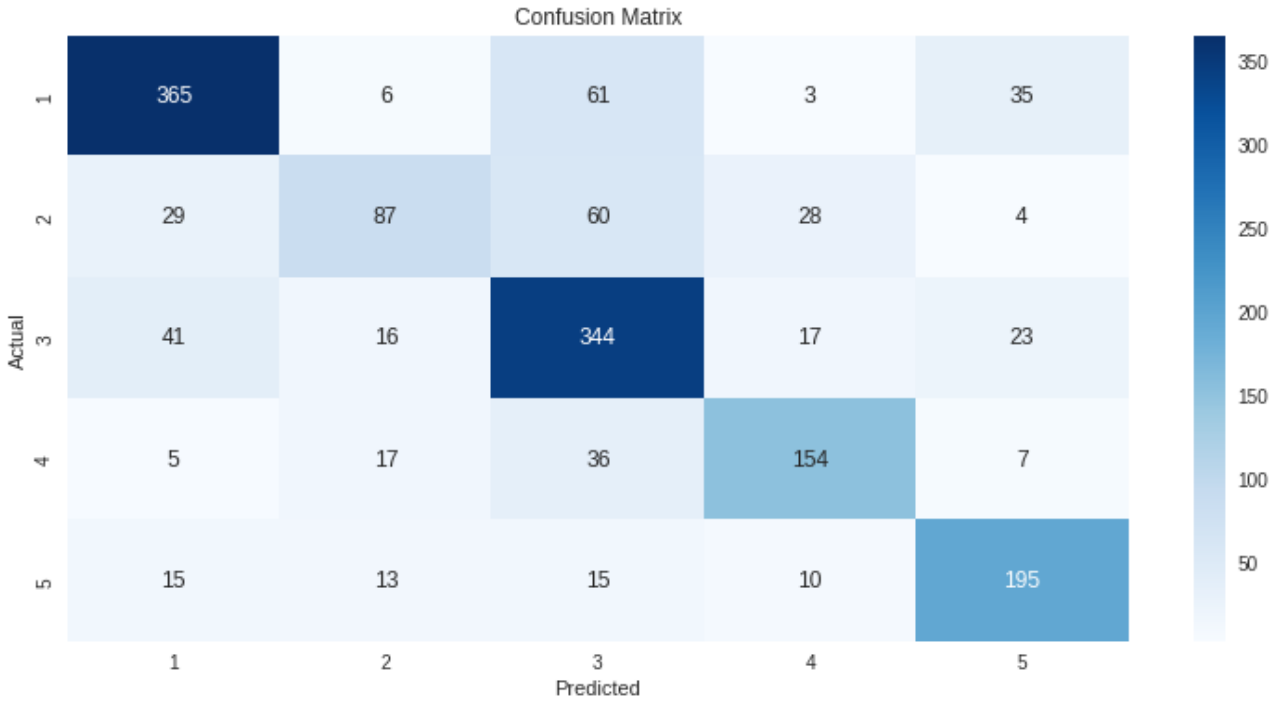

Figure 5-25 Confusion Matrix of XGBoost Model

The ROC curve of the developed XGBooost model is illustrated in Figure 5-26. It can be seen that the AUC of all classes is relatively high, indicating low misclassification among different classes. Macro-average AUC value of 0.92 is close to GBT's one.

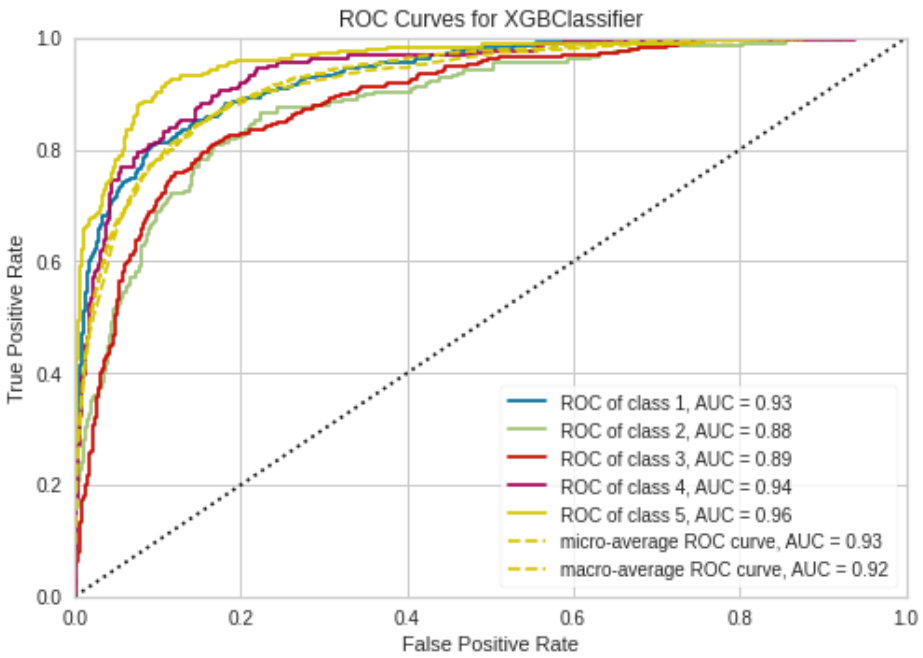

Figure 5-26 ROC Curves for XGBoost Model

The generated XGBoost model's final evaluation metrics are described in Table 5-16 at the end. Similar to the GBT model, satisfactory accomplishments for pipes with condition ratings of 1,3,4 and 5 are visible. Precision value of 0.80 for class 1 and recall value of 0.78 for class 5 are proofs of low misclassification for a crowded group of pipes which are in condition rating of 1 and high ability of the model to capture minority group of pipes which are in condition rating of 5, respectively. The overall F1-score of 0.70 shows an acceptable performance of this model, finally.

Table 5-16 Precision, Recall, and F1-Score Metrics for XGBoost Model

Condition Rating	Precision	Recall	F1 Score
1	0.80	0.78	0.79
2	0.62	0.42	0.50
3	0.67	0.78	0.72
4	0.73	0.70	0.71
5	0.74	0.78	0.76
Macro - Average	*0.71*	*0.70*	*0.70*

5.8 Discussions and Practical Applications

Based on a combined historical inspection dataset gathered from the cities of Tampa and Dallas, this study developed seven different models to predict the condition level of sewer pipes. The dataset was imbalanced and was resampled by the SVM-SMOTE method as described in section 5.4.3. To establish the prediction models, nine independent variables were included: pipe age, material, diameter, length, depth, slope, soil type, soil pH, and pipe location. Sewer pipe condition ratings were the target variable and were evaluated using the PACP method. It was intended to develop a model to anticipate each pipe's five condition ratings. Figure 5-27 shows the hierarchy of seven developed models using various methods.

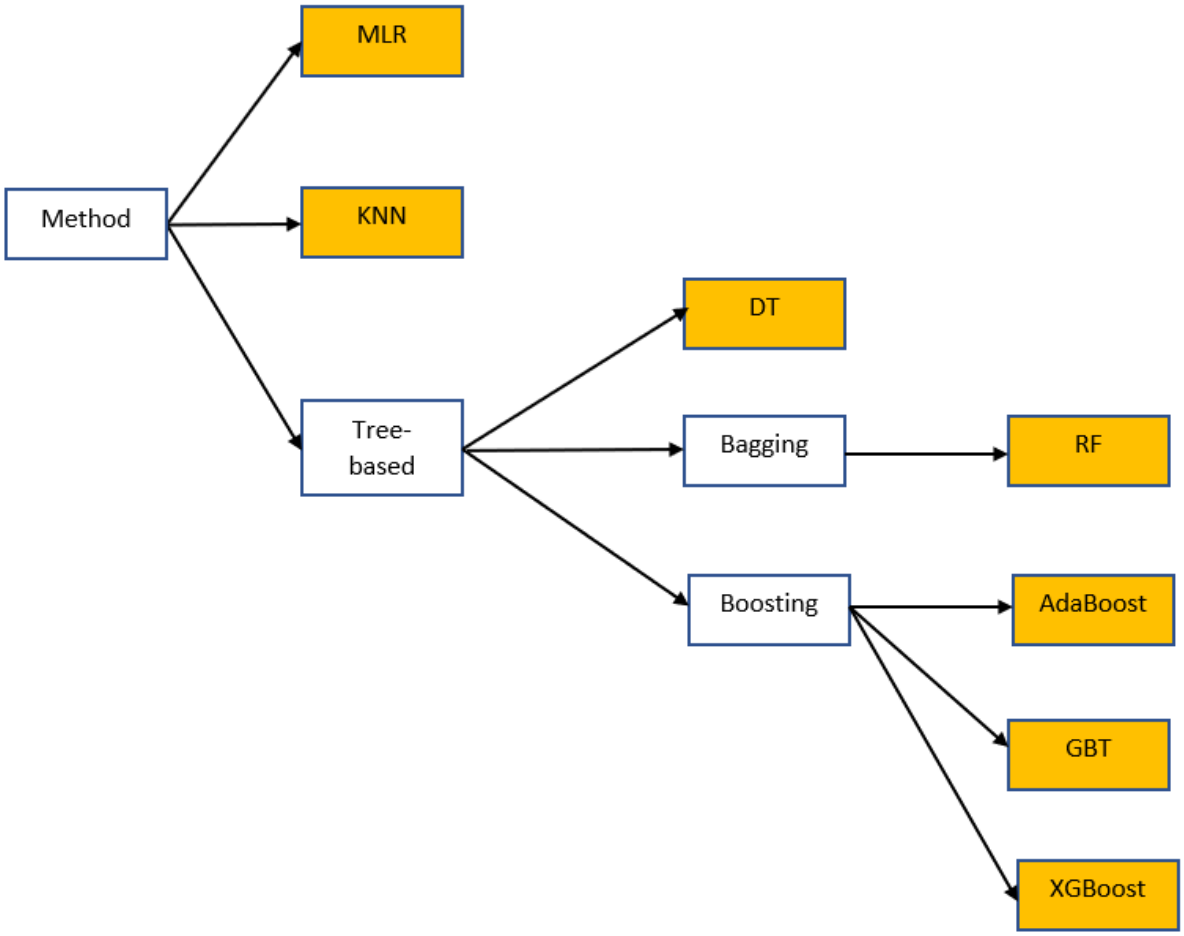

Figure 5-27 Developed Models in this Study

Discussions

Utilizing the confusion matrix, ROC curve, and macro-average F1-score as three different validation techniques, all the models were tested for accuracy. The effectiveness of the models implemented in this investigation is shown in Figure 5-28. Overall F1-score as summary indicator of model's performance is used in this figure. As it can be seen, Random Forest model had the best accuracy, with F1-score of 80%, whereas Multinomial Logistic Regression had the lowest accuracy (F1=48%).

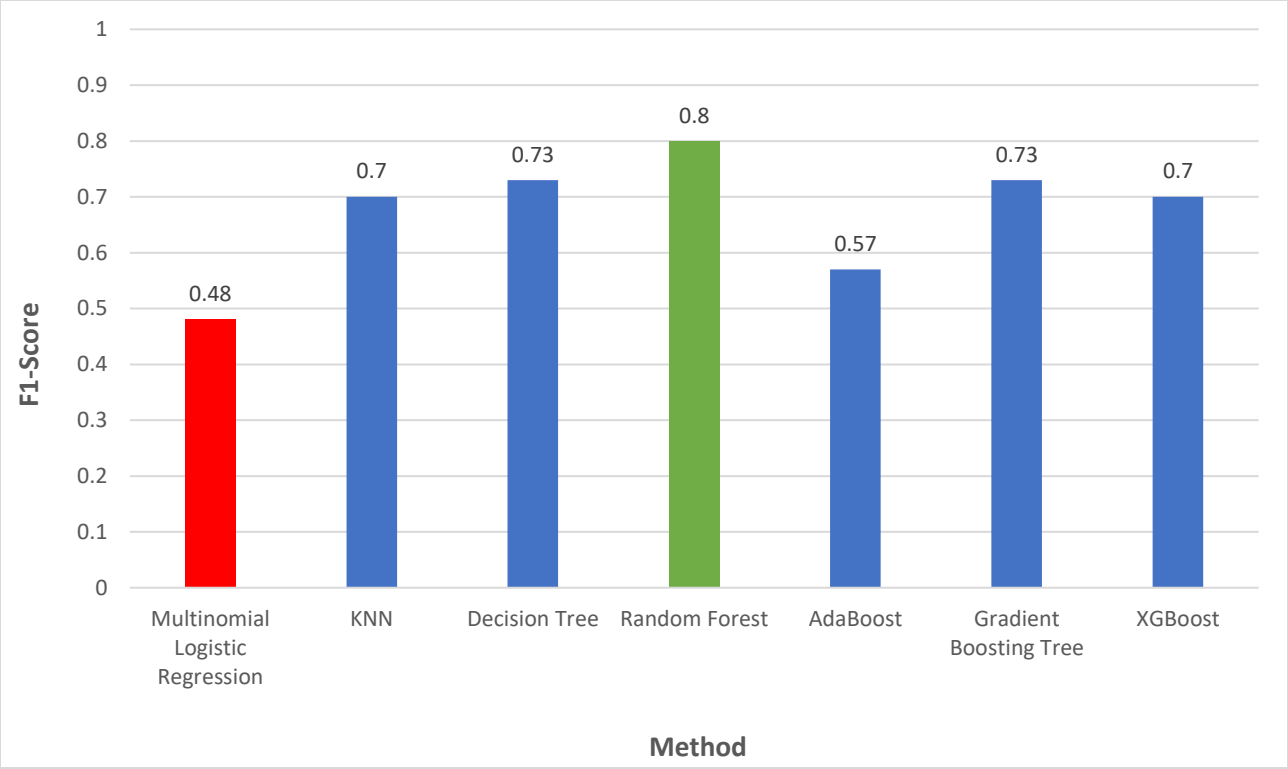

Figure 5-28 Comparison of Model Performances

It can be seen that tree-based models had better performance than others. However, the Bagging approach was more efficient rather than Boosting approach. Furthermore, determining the significant variables is a critical component of condition prediction modeling. These factors have a significant impact on sewer pipe conditions. Therefore, leaving them out of the model could reduce its accuracy. One advantage of tree-based models is the ability to prioritize the significance of the independent variables for both regression and classification goals. Generally, feature importance gives a score indicating how helpful a variable is in putting the model into practice. Figure 5-29 displays the relative importance of various factors obtained from the feature importance attribute of this study's best model (Random Forest).

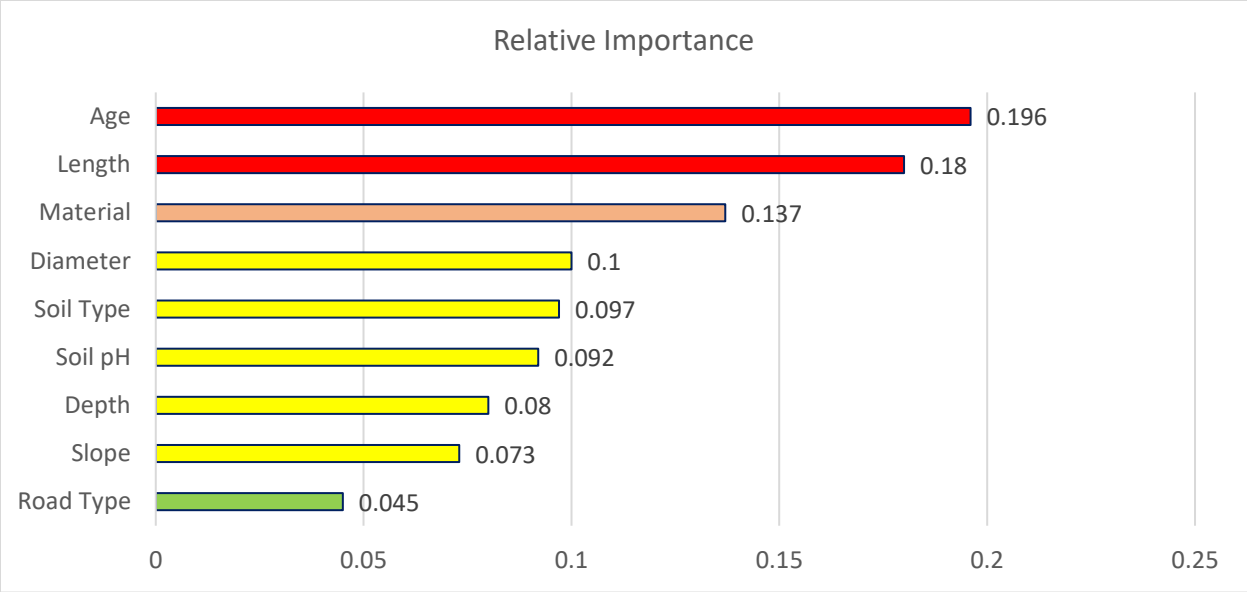

Figure 5-29 Relative Importance of Independent Variables

The figure above demonstrates that factors of age and length of pipes had the highest effect on their condition. The material was the next important parameter. Pipe size surrounded soil type, pH of surrounded soil, depth of buried pipe, and pipe slope had a lower impact, respectively. The pipe location (road type) had the least effect on the condition of sewers.

Practical Applications

The final selected prediction model was used separately for the cleaned dataset of Dallas and Tampa to use it as a support tool in the sewer asset management of both case studies. After running the model for those datasets, the confusion matrix for each one was created, and overall macro-average F1-score were calculated based on the matrixes. Interestingly, the obtained F1 scores were 0.84 and 0.77 for Tampa and Dallas, respectively, which shows good accuracy regarding the accuracy of previously developed models studied in the literature review. Specifically, the F1-score for pipes with a condition rating of 5 was investigated. It was 0.66 for Dallas and 0.83 for Tampa. So, for Dallas city, the suggestion would be to design strategic investment plans in order to know the quantity of pipes that the utility should rehabilitate or replace because the F1-score for the pipes with a condition rating of 5 in this city shows that the developed model has not had high accuracy in detecting them and more inspection data is needed. On the other hand, for Tampa city, the suggestion would be to make decisions concerning rehabilitation plans

instead of more inspection since the accuracy of correctly identifying the sewers in a critical condition was 83%, which is high (Hernandez et al., 2018).

Furthermore, the developed Random Forest model could be utilized as a tool to identify the condition rating of uninspected sewer pipes. By entering the digit of numerical variables like age and length and the class of categorical variables like material type in this model, the model would predict the condition rating of the intended pipe by showing a score between 1 to 5. This result could be used in the decision-making process for replacement or rehabilitation and inspection prioritization too. The mentioned code is available upon request.

5.9 Justification of Results

This study's results were consistent with the results of some studies conducted by other researchers. The accuracy of different developed prediction models and the significant factors affecting the condition of sewer pipes were investigated in the current study. Tables 5-17 and 5-18 show the comparison of results between this study and some similar previous studies.

Table 5-17 Similar Results Regarding Important Parameters Affecting the Condition of Sewer Pipes

Current Study	Other Studies
Age	• Lubini and Fuamba (2011) • Tscheikner-Gratl et al. (2014) • Laakaso et al., (2018) • Malek Mohammadi (2019) • Atambo (2021)
Length	• Ana et al. (2009) • Khan et al. (2010) • Malek Mohammadi (2019) • Atambo (2021)
Material	• Davies et al. (2001) • Lubini and Fuamba (2011) • Malek Mohammadi (2019)

Table 5-18 Comparison of Various Models' Accuracies

Method	Current Study	Other Studies	
MLR	48%	• Salman and Salem (2012) • Laakaso et al., (2018) • Malek Mohammadi (2019) • Loganathan (2021) • Atambo (2021)	52% 62% 65% 44% 75%
KNN	70%	• Malek Mohammadi (2019) • Loganathan (2021)	83% 83%
DT	73%	• Mazumder et. al (2021)	77%
RF	80%	• Laakso et al. (2018) • Hernandez et al. (2018) • Loganathan (2021)	62% 63% 94%
AdaBoost	57%	• Mazumder et. al (2021)	74%
GBT	73%	• Malek Mohammadi (2019)	87%
XGBoost	70%	• Mazumder et. al (2021)	85%

5.10 Chapter Summary

This chapter presented a detailed overview of model development and model results. The resampling process of the dataset was thoroughly explained. The technique of cross-validation to avoid overfitting was described. Firstly, the development and results of Binary Logistic Regression to create deterioration curves were shown. Then, the training and testing process of seven algorithms, including Multinomial Logistic Regression, K-Nearest Neighbors, regular Decision Tree, Random Forest, AdaBoost, Gradient Boosting Tree, and XGBoost, were demonstrated. The confusion matrix, ROC curve, precision, recall, and F1-sore metrics validated their results. Also, by comparing their performance, the best model to predict the condition of sewer pipes was selected, and also, significant factors affecting the condition of pipes were identified. Finally, practical applications and results justification of this study were discussed.

Chapter 6 Conclusions, Limitations, and Recommendations for Future Research

6.1 Conclusions

The following conclusions have been drawn due to the development of the various prediction models investigated in this study. Separate summaries of the development process, each model's outcomes, and comparison findings are provided.

Data Collection and Resampling

- Sewer pipe network datasets of two cities, including Dallas, TX, and Tampa, FL, were combined in this study. The reasons for the combination were increasing the number of data points to increase the model's accuracy, developing a more comprehensive model than previous studies, and avoiding the model's overfitting.
- The resampling was done since the available dataset was imbalanced, resulting in the poor performance of the models. Various resampling approaches were examined, and SVM-SMOTE was found to be the most useful one.

Binary Logistic Regression

- For the test dataset, Binary Logistic Regression obtained an overall correct prediction percentage of 88%. The area under the ROC curve was 0.73, which was satisfactory.
- According to results of Binary Logistic Regression, PVC pipes deteriorate slower than other pipe materials. Also, longer pipes are more prone to degradation. In addition, pipes buried in clay have the higher deterioration rate than pipes surrounded by other soils.

Multinomial Logistic Regression

- F1-scores of 0.29 and 0.15 for pipes with condition ratings of 2 and 4, respectively, indicate the weak performance of the developed Multinomial Logistic Regression model in predicting these classes. The overall macro-average F1-score of 0.48 states that the developed MLR model was not a reliable model.

KNN

- An overall macro-average F1-score of 0.70 shows an acceptable performance of the developed KNN model for the available sewer pipes dataset. It was found that the KNN model had better performance for pipes with a condition rating of 5 (F1= 0.76) rather than

others and had the lowest accuracy in predicting pipes with a condition rating of 2 (F1=0.60).

Decision Tree

- F1-scores of more than 0.7 for all classes showed that the developed Decision Tree was a suitable model for pipes in all five condition ratings. Overall, the Decision Tree classifier's model, which had a macro-average F1-score of 0.73, outperformed the KNN model.

Random Forest

- Random Forest was a tree-based model with a bagging approach developed in this study. Relatively small values in the non-diagonal cells of its confusion matrix demonstrated that this model had a low misclassification rate. AUC for ROC curves of all condition ratings was close to 1, showing the high efficiency of the model. Finally, the overall macro-average F1-score of 0.8 showed that the model had an outstanding performance in predicting all classes and outperformed the KNN and Decision Tree models.

AdaBoost

- AdaBoost was the first tree-based algorithm with a Boosting approach to be tried. The values in non-diagonal cells of its confusion matrix were high, illustrating relatively high misclassification among different classes, specifically between classes 3 and 4. The precision and recall values of the model for pipes with a condition rating of 5, which has always been a sign of the model's performance, were low. Lastly, the overall macro-average F1-score of 0.57 indicates that AdaBoost was not a suitable model for the available dataset. The interesting point was that AdaBoost was the only tree-based model in this research with weak performance.

Gradient Boosting Tree

- Another Boosting approach tried in this study was Gradient Boosting Tree. The values on diagonal elements disclosed a relatively suitable result for the developed model, specifically for pipes with condition ratings of 1 and 3. AUC values for all classes were higher than 0.9, which indicated that the model had a much more accuracy than the

previous boosting model. The calculated overall F1-score of 0.73 showed that this model had good performance.

XGBoost

- Last developed prediction model by machine learning methods was XGBoost model. Macro-average AUC value of 0.92 was close to GBT's one. Precision value of 0.80 for class 1 and recall value of 0.78 for class 5 were proofs of low misclassification for a crowded group of pipes which are in condition rating of 1 and high ability of the model to capture minority group of pipes which are in condition rating of 5, respectively. The overall F1-score of 0.70 showed an acceptable performance of this model, finally.

Comparison of Models

- Overall macro-average F1-score as a summary indicator of the model's performance was used to compare the developed models. The Random Forest model had the best accuracy in predicting the condition rating of pipes in all five conditions, with an F1-score of 0.8. In contrast, Multinomial Logistic Regression had the lowest accuracy with an F1-score of 0.48. It could be seen that tree-based models had better performance than others. However, the Bagging approach was more efficient rather than Boosting approach.
- Also, results of most accurate model developed in this dissertation (Random Forest) demonstrated that factors of age, length of pipes and pipe material had the highest effect on the condition ratings of sewers, which was consistent with results of Binary Logistic Regression model except surrounded soil type. On the other hand, pipe location (road type) had the least effect.

6.2 Limitations of this Research

The lack of a suitable dataset to create the models is the fundamental limitation of condition prediction modeling. As was previously mentioned, this dissertation's use of a merged dataset from two cities provided some advantages. However, one of the drawbacks is that certain environmental elements, such as the type of flow in pipes, which differs in each geographic location due to varying viscosity and population in different urban areas, would be neglected when the dataset was combined. Also, in the

dataset, sewer pipe segments' lengths were measured from manhole to manhole. The other important limitation of this study was the lack of information regarding the number of joints in each pipe section.

6.3 Recommendations for Future Research

Additional research studies could be accomplished to further improve the research work discussed in this dissertation. Potential future development could include but not limited to the following:

- In this study, various resampling methods were discussed. Potential research could be manipulating the dataset by different resampling methods, developing the models by diverse machine learning procedures, and validating the efficiency of resampling methods using the testing results of each machine learning procedure. In this study, this process was done, but it has much more potential for further investigations.
- Studies might involve sewer pipe segments with a maintenance activity history.
- The number of joints in the inspected sewer pipe segment could be taken into account when developing the model.
- The developed models of this study could be improved by including wastewater type and volumetric flow rate variables.
- Consequences of failure could be considered in order to do the risk analysis.
- Combining the dataset of more cities especially with various slopes, could result in more comprehensive prediction model.
- In machine learning, hyperparameter optimization or tuning is choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value controls the learning process. The potential research could be to investigate various tuning processes in order to increase the efficiency of sewer condition prediction models.

References

Agresti, A. (2018). *An introduction to categorical data analysis*. John Wiley & Sons.

Ariaratnam, S. T., El-Assaly, A., & Yang, Y. (2001). Assessment of infrastructure inspection needs using logistic models. *Journal of infrastructure systems*, *7*(4), 160-165.

ASCE (2020). Report Card for America's Infrastructure. New York: American Society of Civil Engineers (ASCE).

Atambo, D. O. (2021). Development and comparison of prediction models for sanitary sewer pipes condition assessment using multinomial logistic regression and artificial neural network (Order No. 28826488). Available from Dissertations & Theses @ University of Texas - Arlington; ProQuest Dissertations & Theses Global. (2596630654). Retrieved from https://login.ezproxy.uta.edu/login.

Atambo, D. O., Najafi, M., & Kaushal, V. (2022). Development and Comparison of Prediction Models for Sanitary Sewer Pipes Condition Assessment Using Multinomial Logistic Regression and Artificial Neural Network. Sustainability, 14(9), 5549.

Bakry, I., Alzraiee, H., Kaddoura, K., El Masry, M., & Zayed, T. (2016). Condition prediction for chemical grouting rehabilitation of sewer networks. *J. Perform. Constr. Facil*, *30*(6), 04016042.

Bakry, I., Alzraiee, H., Masry, M. E., Kaddoura, K., & Zayed, T. (2016). Condition prediction for cured-in-place pipe rehabilitation of sewer mains. *J. Perform. Constr. Facil*, *30*(5), 04016016.

Barqawi, A., & Ahmad, H. (2006). *Condition rating models for underground infrastructure: Sustainable water mains* (Doctoral dissertation, Concordia University).

Behboudian, J., & Asgharzadeh, A. (2008). On the distribution of z-scores (Research Note).

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197-227.

Bishop, C. M., & Nasrabadi, N. M. (2016). Pattern Recognition and Machine Learning, vol. 4, no. 4.

Burian, S. J., Nix, S. J., Pitt, R. E., & Durrans, S. R. (2000). Urban wastewater management in the United States: Past, present, and future. *Journal of Urban Technology*, *7*(3), 33-62.

Charniak, E., & McDermott, D. (1985). Introduction to Artificial. *Intelligence (Addison Wesley, Reading, MA, 1984)*.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Chughtai, F., & Zayed, T. (2008). Infrastructure condition prediction models for sustainable sewer pipelines. *Journal of Performance of Constructed Facilities*, *22*(5), 333-341.

Coles, S., Bawa, J., Trenner, L., & Dorazio, P. (2001). *An introduction to statistical modeling of extreme values* (Vol. 208, p. 208). London: Springer.

Dasu, T., & Johnson, T. (2003). *Exploratory data mining and data cleaning*. John Wiley & Sons.

Davies, J. P., Clarke, B. A., Whiter, J. T., & Cunningham, R. J. (2001). Factors influencing the structural deterioration and collapse of rigid sewer pipes. *Urban water*, *3*(1-2), 73-89.

E. Ana, W. Bauwens, M. Pessemier, C. Thoeye, S. Smolders, I. Boonen & G. De Gueldre. (2009). An investigation of the factors influencing sewer structural deterioration. Urban Water Journal, 6:4, 303-312, DOI: 10.1080/15730620902810902.

Elnahas, M. M., Hussein, M., & Keshk, A. (2022). Imbalanced Data Oversampling Technique Based on Convex Combination Method. *IJCI. International Journal of Computers and Information*, *9*(1), 15-28.

EPA, (2004), Asset Management for Sewer Collection Systems. Washington, DC: Office of Wastewater Management.

EPA, (2010), Report on Condition Assessment of Wastewater Collection Systems. July 2018, <www. Ep.gov/nrmrl>.

EPA, (2015), Condition Assessment of Underground Pipes. April 2018, <www.Ep.gov/nrmrl>.

Federal Highway Administration (FHWA). (2011). Washington, D.C.

Flintsch, G. W., & Chen, C. (2004). Soft computing applications in infrastructure management. *Journal of Infrastructure Systems*, *10*(4), 157-166.

Gedam, A., Mangulkar, S., & Gandhi, B. (2016). Prediction of sewer pipe main condition using the linear regression approach. *Journal of geoscience and environment protection*, *4*(5), 100-105.

Géron, A. (2017). Hands-on machine learning with scikit-learn and tensorflow: Concepts. *Tools, and Techniques to build intelligent systems*.

Ghorbani, R., & Ghousi, R. (2020). Comparing different resampling methods in predicting students' performance using machine learning techniques. *IEEE Access*, *8*, 67899-67911.

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*.

Guzmán-Fierro, J., Charry, S., González, I., Peña-Heredia, F., Hernández, N., Luna-Acosta, A., & Torres, A. (2020). Bayesian network-based methodology for selecting a cost-effective sewer asset management model. *Water Science and Technology*, *81*(11), 2422-2431.

Harvey, R. R., & McBean, E. A. (2014). Comparing the utility of decision trees and support vector machines when planning inspections of linear sewer infrastructure. *Journal of Hydroinformatics*, *16*(6), 1265-1279.

Harvey, R. R., & McBean, E. A. (2014). Predicting the structural condition of individual sanitary sewer pipes with random forests. *Canadian Journal of Civil Engineering*, *41*(4), 294-303.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.

Hawari, A., Alkadour, F., Elmasry, M., & Zayed, T. (2017). Simulation-based condition assessment model for sewer pipelines. *Journal of Performance of Constructed Facilities*, *31*(1), 04016066.

Hernández, N., Caradot, N., Sonnenberg, H., Rouault, P., & Torres, A. (2018). Support tools to predict the critical structural condition of uninspected pipes for case studies of Germany and Colombia. *Water Practice & Technology*, *13*(4), 794-802.

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). Applied logistic regression. Wiley, Hoboken, NJ.

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, *5*(2), 1.

Hou, Y., Lei, D., Li, S., Yang, W., & Li, C. Q. (2016). Experimental investigation on corrosion effect on mechanical properties of buried metal pipes. *International Journal of Corrosion*, *2016*.

Jeong, H. S., Baik, H. S., & Abraham, D. M. (2005). An ordered probit model approach for developing markov chain based deterioration model for wastewater infrastructure systems. In *Pipelines 2005: Optimizing Pipeline Design, Operations, and Maintenance in Today's Economy* (pp. 649-661).

Kabir, G., Balek, N. B. C., & Tesfamariam, S. (2018). Sewer structural condition prediction integrating Bayesian model averaging with logistic regression. *Journal of Performance of Constructed Facilities*, *32*(3), 04018019.

Kaushal, V., & Guleria, S. P. (2015). Geotechnical investigation of black cotton soils. *International Journal of Advances in Engineering Sciences*, *5*(2), 15-22.

Khan, Z., Zayed, T., & Moselhi, O. (2010). Structural condition assessment of sewer pipelines. *Journal of performance of constructed facilities*, *24*(2), 170-179.

Kleiner, Y., Sadiq, R., & Rajani, B. (2004, August). Modeling failure risk in buried pipes using fuzzy Markov deterioration process. In *ASCE international conference on pipeline engineering and construction* (pp. 1-12).

Kley, G., and Caradot, N. (2013). D1.2. Review of sewer deterioration models. KWB report, project SEMA, Berlin, Germany.

Kulandaivel, G. (2004). *Sewer pipeline condition prediction using neural network models*. Michigan State University.

Laakso, T., Kokkonen, T., Mellin, I., & Vahala, R. (2018). Sewer condition prediction and analysis of explanatory factors. *Water*, *10*(9), 1239.

Loganathan, K. (2021). *Development of a Model to Prioritize Inspection and Condition Assessment of Gravity Sanitary Sewer Systems* (Doctoral dissertation, The University of Texas at Arlington).

Loganathan, K., Najafi, M., Kaushal, V., & Covilakam, M. (2022). Development of a Decision Support Tool for Inspection and Monitoring of Large-Diameter Steel and Prestressed Concrete Cylinder Water Pipes. *Journal of Pipeline Systems Engineering and Practice*, *13*(1), 04021067.

Loh, W. Y. (2014). Fifty years of classification and regression trees. *International Statistical Review*, *82*(3), 329-348.

Lubini, A. T., & Fuamba, M. (2011). Modeling of the deterioration timeline of sewer systems. *Canadian Journal of Civil Engineering*, *38*(12), 1381-1390.

Luger, G. F. (2009). Artificial Intelligence. Structures and Strategies for Complex Problem Solving. Addison Wesley, Boston MA.

Malek Mohammadi, M. (2019). Development of Condition Prediction Models for Sanitary Sewer Pipes. Doctoral Dissertation. University of Texas at Arlington. Arlington, TX, USA.

Malek Mohammadi, M., Najafi, M., Kaushal, V., Serajiantehrani, R., Salehabadi, N., & Ashoori, T. (2019). Sewer pipes condition prediction models: a state-of-the-art review. *Infrastructures*, *4*(4), 64.

Malek Mohammadi, M., Najafi, M., Kermanshachi, S., Kaushal, V., & Serajiantehrani, R. (2020). Factors influencing the condition of sewer pipes: State-of-the-art review. *Journal of Pipeline Systems Engineering and Practice*, *11*(4), 03120002.

Mashford, J., Marlow, D., Tran, D., & May, R. (2011). Prediction of sewer condition grade using support vector machines. *Journal of Computing in Civil Engineering*, *25*(4), 283-290.

Mazumder, R. K., Salman, A. M., & Li, Y. (2021). Failure risk analysis of pipelines using data-driven machine learning algorithms. *Structural safety*, *89*, 102047.

McDonald, J. H. (2009). *Handbook of biological statistics* (Vol. 2, pp. 6-59). Baltimore, MD: sparky house publishing.

Melosi, M. V. (2000). *The sanitary city: Urban infrastructure in America from colonial times to the present* (p. 107). Baltimore: Johns Hopkins University Press.

Micevski, T., Kuczera, G., & Coombes, P. (2002). Markov model for storm water pipe deterioration. *Journal of infrastructure systems*, *8*(2), 49-56.

Misiunas, D. (2005). *Monitoring and asset condition assessment in water supply systems* (Doctoral dissertation, PhD thesis, Lund Univ., Lund).

Mohammadi, M. M., Najafi, M., Tabesh, A., Riley, J., & Gruber, J. (2019). Condition prediction of sanitary sewer pipes. *Pipelines*, *2019*, 117-126.

Moteleb, M. (2010). *Risk based decision making tools for sewer infrastructure management*. University of Cincinnati.

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: a guide for data scientists*. " O'Reilly Media, Inc."

Najafi, M., & Kulandaivel, G. (2005). Pipeline condition prediction using neural network models. In Pipelines 2005: Optimizing Pipeline Design, Operations, and Maintenance in Today's Economy (pp. 767-781).

Najafi, M., Gokhale, S., Calderón, D. R., & Ma, B. (2022). *Trenchless technology: Pipeline and utility design, construction, and renewal*. McGraw-Hill Education.

National Association of Sewer Service Companies (NASSCO) (2018), Pipeline Assessment Certification Manual. Marriottsville, MD, USA, March 2018.

Opila, M. C. (2011). *Structural condition scoring of buried sewer pipes for risk-based decision making*. University of Delaware.

O'REILLY, M. P., Rosbrook, R. B., Cox, G. C., & McCloskey, A. (1989). *Analysis of defects in 180km of pipe sewers in Southern Water Authority* (No. RR 172).

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), 154-164.

Rawlings, J. O. (1989). Applied regression analysis: a research tool. Wadsworth & Brooks. *Pacific Grove, CA*.

Salman, B. (2010). *Infrastructure management and deterioration risk assessment of wastewater collection systems*. University of Cincinnati.

Salman, B., & Salem, O. (2012). Modeling failure of wastewater collection lines using various section-level regression models. Journal of Infrastructure Systems, 18(2), 146-154.

Simon, P. (2013). *Too big to ignore: the business case for big data* (Vol. 72). John Wiley & Sons.

Singh, A., & Adachi, S. (2013). Bathtub curves and pipe prioritization based on failure rate. *Built Environment Project and Asset Management*.

Sousa, V., Matos, J. P., & Matias, N. (2014). Evaluation of artificial intelligence tool performance and uncertainty for predicting sewer structural condition. *Automation in Construction*, *44*, 84-91.

Syachrani, S. (2010). *Advanced sewer asset management using dynamic deterioration models* (Doctoral dissertation, Oklahoma State University).

Syachrani, S., Jeong, H. S. D., & Chung, C. S. (2013). Decision tree–based deterioration model for buried wastewater pipelines. *Journal of Performance of Constructed Facilities*, *27*(5), 633-645.

Tafuri, A. N., & Dzuray, E. J. (2000). Sewer pipeline performance indicators: Learning from the European experience. In *Building Partnerships* (pp. 1-10).

Taneja, S., Suri, B., & Kothari, C. (2019, September). Application of balancing techniques with ensemble approach for credit card fraud detection. In *2019 International Conference on Computing, Power and Communication Technologies (GUCON)* (pp. 753-758). IEEE.

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big Data*, *7*(1), 1-47.

Teh, K., Armitage, P., Tesfaye, S., Selvarajah, D., & Wilkinson, I. D. (2020). Imbalanced learning: Improving classification of diabetic neuropathy from magnetic resonance imaging. *PloS one*, *15*(12), e0243907.

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.

Tran, D. H., Ng, A. W. M., Perera, B. J. C., Burn, S., & Davis, P. (2006). Application of probabilistic neural networks in modelling structural deterioration of stormwater pipes. *Urban Water Journal*, *3*(3), 175-184.

Wasim, M., Shoaib, S., Mubarak, N. M., & Asiri, A. M. (2018). Factors influencing corrosion of metal pipes in soils. *Environmental Chemistry Letters*, *16*(3), 861-879.

Wirahadikusumah, R., Abraham, D., & Iseley, T. (2001). Challenging issues in modeling deterioration of combined sewers. *Journal of infrastructure systems*, *7*(2), 77-84.

Yijing, L., Haixiang, G., Xiao, L., Yanan, L., & Jinling, L. (2016). Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. *Knowledge-Based Systems*, *94*, 88-104.

Zheng, X. (2020). *SMOTE variants for imbalanced binary classification: heart disease prediction*. University of California, Los Angeles.

Appendix A

Abbreviations

-2LL - -2 Log likelihood

AC - asbestos-cement

Adaboost - Adaptive Boosting

AI – Artificial Intelligence

ANN - Artificial Neural Network

ASCE – American Society of Civil Engineers

AUC – Area under the Curve

AWWA – American Water Works Association

BPNN - Backpropagation Neural Network

CBD - central business district

CCTV - Closed-circuit television

CF – Clear Fork

CI – Cast Iron

CIPP – Cured-in-place Pipe

COF – Consequence of Failure

CSS – Combined Sewer Systems

CUIRE - Center for Underground Infrastructure Research and Education

DFW - Dallas Fort Worth

DI - ductile iron

DT – Decision Trees

DWU - Dallas Water Utilities

EPA – Environmental Protection Agency

FHWA – Federal Highway Administration

FN – False Negative

FNR - False Negative Rate

FP – False Positive

FPR – False Positive Rate

FRP -Fiberglass Reinforced Plastic

ft - Feet

GBT - Gradient Boosting Tree

Gi – Gini Index

GIS - Geographical Information System

GPS – Global Positioning System

HDPE - High Density Polyethylene

in. - Inch

k-NN – k-Nearest Neighbors

LOF – Likelihood of Failure

LR – Logistic Regression

MCMC – Markov Chain Monte Carlo

MDA – Mean Decrease Accuracy

MDI – Mean Decrease Impurity

MLE – Maximum Likelihood Estimation

MLR - Multinomial Logistic Regression

MSE - Mean Squared Error

NASSCO - National Association of Sewer Service Companies

NRC – National Research Council Canada

O&M - Operation and maintenance

OMR – Operational Maintenance Rating

OR – Odds Ratio

OR – Overall Pipe Rating

OSR - Operational Structural Rating

PACP - Pipeline Assessment and Certificate Program

PCCP - prestressed concrete cylinder pipe

PCCP – Prestressed Concrete Cylinder Pipe

PCP - polymer concrete pipe

POF - Probability of failure

PVC - Polyvinyl Chloride Pipe

RCP – Reinforced Concrete Pipe

RF - Random Forest

RI – Rating Index

RMSE - root mean squared error

ROC – Receiver Operator Characteristic

SBS – Sequential Backward Selection

SE – Standard Error

SFFS – Sequential Forward Floating Selection

SFS – Sequential Forward Selection

SPSS – IBM SPSS Statistics Packages

SSS – Separate Sanitary Sewer and Storm Sewer System

SVMs - Support Vector Machines

TN – True Negative

TN – True Negative

TP – True Positive

TPR – True Positive Rate

U.S. – United States

USEPA - United States Environmental Protection Agency

UTA - The University of Texas at Arlington

VCP - Vitrified clay pipe

WRc – Water Research Center

WSAA – Water Services Association of Australia

XGBoost - Extreme Gradient Boosting

Appendix B

Data Sample (1,000 pipe segments)

If you need more data and/or Python code of developed models, please contact:

salar.shirkhanloo@mavs.uta.edu.

N	Age	Material	Diameter	Slope	Depth	Length	Soil PH	Soil Type	Pipe Location	Condition Rating
1	69	RC	27	0.14	5.00	2261.24	7.9	Clay	Street	1
2	73	RC	54	0.16	6.00	2054.15	8.2	Loam	Alley	1
3	53	RC	21	0.03	6.00	1901.69	7.9	Clay	Street	3
4	18	RC	66	0.03	7.00	1844.36	7.9	Clay	Alley	1
5	55	RC	30	0.16	6.00	1771.92	8.2	Loam	Street	1
6	64	RC	30	0.16	6.00	1746.00	8.2	Loam	Street	3
7	74	RC	60	0.10	6.00	1714.24	7.9	Clay	Street	3
8	74	RC	90	0.07	10.00	1681.10	7.9	Clay	Street	5
9	44	RC	84	0.03	7.00	1634.70	7.9	Clay	Street	1
10	75	RC	72	0.08	10.00	1586.91	7.9	Clay	Street	1
11	18	RC	66	0.03	8.00	1584.52	7.9	Clay	Alley	1
12	64	RC	30	0.16	6.00	1533.89	8.2	Loam	Street	3
13	65	RC	54	0.20	10.00	1517.08	8.2	Loam	Street	5
14	72	RC	48	0.00	6.00	1484.47	8.2	Loam	Street	3
15	78	RC	24	0.55	7.00	1412.87	8.2	Loam	Highway	5
16	67	RC	30	0.50	6.00	1407.34	7.9	Clay	Street	3
17	57	VCP	8	0.30	12.00	1377.37	7.9	Clay	Street	1
18	67	RC	27	0.80	7.00	1288.34	8.2	Loam	Street	3
19	66	RC	8	0.48	7.00	1256.95	7.9	Clay	Street	3
20	61	RC	10	0.02	8.00	1251.44	8.2	Rock	Alley	1
21	73	RC	24	0.35	10.00	1251.28	8.2	Clay	Street	1
22	56	VCP	8	0.40	7.00	1227.89	7.9	Clay	Street	2
23	71	RC	36	0.06	7.00	1221.63	8.2	Loam	Street	1
24	66	RC	30	0.08	5.00	1216.43	8.2	Loam	Street	5
25	66	RC	8	1.12	5.00	1214.49	8.2	Clay	Street	1
26	48	RC	30	0.18	11.00	1208.03	8.2	Loam	Alley	1
27	49	RC	60	0.16	13.00	1204.24	8.2	Loam	Alley	3
28	65	RC	30	0.20	6.00	1195.22	8.2	Loam	Street	3
29	67	RC	30	0.50	6.00	1192.53	7.9	Clay	Street	3
30	61	VCP	24	0.03	12.00	1192.50	7.9	Clay	Highway	2
31	66	RC	30	0.01	7.00	1182.37	8.2	Loam	Street	3
32	74	RC	90	0.00	10.00	1166.03	7.9	Clay	Street	5
33	70	VCP	8	0.03	7.00	1164.51	8.2	Clay	Street	1
34	68	RC	33	0.46	10.00	1161.29	8.2	Loam	Street	1
35	54	VCP	21	0.32	15.00	1153.63	7.9	Clay	Street	1
36	65	RC	54	0.20	8.00	1150.44	8.2	Loam	Street	3
37	52	RC	30	0.01	6.00	1149.54	7.9	Clay	Street	3

38	50	VCP	8	0.30	7.00	1148.24	6.8	Sand	Street	1
39	50	RC	6	0.01	7.00	1132.55	8.2	Loam	Street	3
40	50	RC	60	0.18	13.00	1124.80	8.2	Loam	Alley	3
41	64	RC	36	0.10	7.00	1119.67	8.2	Loam	Highway	3
42	65	VCP	8	0.30	7.00	1104.19	7.9	Clay	Easement	3
43	49	VCP	8	0.32	6.00	1089.25	5.8	Sand	Street	3
44	64	RC	24	0.12	10.00	1072.68	7.5	Clay	Street	3
45	54	RC	30	0.34	5.00	1070.55	8.2	Loam	Highway	3
46	25	RC	72	0.04	8.00	1069.34	7.9	Clay	Highway	1
47	73	RC	72	0.06	10.00	1064.00	7.9	Clay	Street	5
48	38	RC	78	0.11	13.00	1059.28	7.9	Clay	Street	1
49	75	RC	72	0.08	5.00	1055.30	7.9	Clay	Alley	5
50	38	RC	72	0.14	13.00	1053.40	7.9	Clay	Alley	5
51	61	RC	8	0.60	10.00	1050.04	7.9	Clay	Easement	3
52	65	RC	12	0.01	5.00	1042.53	8.2	Clay	Easement	3
53	63	VCP	15	0.20	5.00	1041.29	6.8	Sand	Alley	1
54	63	RC	12	0.20	8.00	1038.69	8.2	Loam	Street	2
55	50	VCP	15	0.20	11.00	1037.16	7.5	Loam	Street	1
56	54	VCP	12	0.01	12.00	1036.35	6.8	Sand	Street	2
57	49	RC	48	0.07	13.00	1032.81	8.2	Loam	Street	5
58	90	RC	42	0.10	5.00	1032.78	7.5	Clay	Highway	3
59	67	RC	39	0.28	20.00	1029.78	8.2	Loam	Alley	1
60	48	RC	30	0.18	11.00	1029.26	8.2	Loam	Street	3
61	36	RC	42	0.15	7.00	1024.02	7.9	Clay	Street	3
62	73	RC	54	0.16	6.00	1022.20	8.2	Loam	Street	3
63	67	RC	30	0.40	6.00	1016.10	7.9	Clay	Street	3
64	38	RC	78	0.11	13.00	1006.47	7.9	Clay	Alley	1
65	64	VCP	8	1.00	12.00	1002.05	7.9	Clay	Highway	1
66	60	VCP	10	0.20	15.00	998.52	7.5	Clay	Street	5
67	40	RC	60	0.03	10.00	997.39	7.5	Clay	Street	3
68	59	VCP	6	0.45	7.00	988.27	6.8	Sand	Street	2
69	41	RC	48	0.16	13.00	985.22	8.2	Loam	Street	1
70	38	RC	72	0.11	13.00	982.95	7.9	Clay	Street	1
71	53	VCP	6	1.60	7.00	981.42	8.2	Clay	Street	2
72	35	PVC	18	0.20	9.00	980.66	7.9	Clay	Alley	1
73	63	VCP	8	0.30	8.00	978.78	7.9	Clay	Alley	3
74	40	RC	54	0.03	10.00	973.87	7.5	Clay	Street	5
75	71	VCP	12	0.14	7.00	969.66	7.9	Clay	Street	5
76	67	VCP	24	0.03	5.00	965.51	7.9	Clay	Street	4
77	60	VCP	6	0.16	7.00	962.47	7.5	Clay	Street	4
78	71	RC	15	0.00	6.00	954.78	6.7	Sand	Street	3

79	62	RC	30	0.07	5.00	954.07	8.2	Loam	Alley	1
80	58	VCP	10	0.40	11.00	952.08	7.9	Clay	Alley	5
81	68	RC	60	0.18	10.00	950.46	8.2	Loam	Street	3
82	18	RC	66	0.03	7.00	943.50	7.9	Clay	Street	1
83	37	RC	72	0.11	13.00	942.47	8.2	Loam	Street	1
84	37	RC	72	0.00	13.00	940.02	8.2	Loam	Street	1
85	58	VCP	8	0.40	5.00	937.11	6.8	Sand	Street	2
86	90	RC	42	0.10	5.00	935.18	6.7	Sand	Street	3
87	57	RC	42	0.14	8.00	930.82	8.2	Loam	Alley	5
88	63	RC	24	0.22	15.00	930.06	7.9	Clay	Street	5
89	67	VCP	10	0.60	7.00	928.68	7.9	Clay	Street	5
90	34	RC	48	0.08	10.00	917.06	7.5	Clay	Street	3
91	35	RC	48	0.08	10.00	916.42	7.5	Clay	Street	1
92	61	VCP	6	2.10	10.00	907.41	8.2	Clay	Street	5
93	37	RC	48	0.08	13.00	901.93	8.2	Loam	Street	5
94	61	VCP	12	0.02	5.00	892.76	8.2	Rock	Street	1
95	68	RC	42	0.00	10.00	890.39	8.2	Loam	Street	3
96	62	RC	8	0.01	10.00	890.22	8.2	Loam	Alley	4
97	38	RC	72	0.14	13.00	889.25	7.9	Clay	Street	1
98	78	VCP	6	1.64	8.00	889.19	8.2	Clay	Street	3
99	56	VCP	21	0.26	6.00	887.04	6.7	Sand	Alley	2
100	39	PVC	15	0.10	10.00	886.35	7.9	Clay	Street	1
101	57	VCP	6	0.12	12.00	872.95	7.9	Clay	Street	3
102	45	RC	54	0.14	5.00	871.57	6.7	Sand	Street	5
103	53	VCP	8	0.20	7.00	860.05	6.8	Sand	Street	3
104	64	RC	10	1.40	10.00	858.75	7.9	Clay	Easement	3
105	55	VCP	8	0.20	7.00	856.06	6.8	Sand	Street	2
106	29	PVC	8	0.01	8.00	849.35	7.9	Clay	Street	1
107	66	RC	27	0.16	15.00	844.40	7.9	Clay	Street	5
108	57	VCP	15	0.10	5.00	843.43	7.9	Clay	Street	1
109	47	VCP	15	0.01	8.00	838.65	8.2	Loam	Street	2
110	33	PVC	24	0.30	10.00	829.94	8.2	Clay	Street	2
111	73	RC	27	0.64	5.00	825.27	8.2	Clay	Street	5
112	84	RC	36	0.42	7.00	823.92	7.9	Clay	Street	3
113	64	VCP	15	0.28	7.00	823.28	7.5	Clay	Street	2
114	65	RC	30	0.20	6.00	821.42	8.2	Loam	Street	3
115	65	RC	54	0.30	10.00	816.23	8.2	Loam	Alley	3
116	50	VCP	10	2.00	7.00	815.00	8.2	Rock	Alley	1
117	58	VCP	8	0.01	5.00	812.95	7.9	Clay	Street	2
118	77	RC	18	0.84	10.00	812.32	8.2	Loam	Alley	1
119	63	RC	27	0.16	15.00	810.03	7.9	Clay	Street	3

120	64	VCP	8	0.12	8.00	807.70	7.9	Clay	Street	5
121	64	RC	24	0.12	10.00	805.38	6.7	Sand	Street	3
122	42	PVC	15	0.01	10.00	802.72	6.5	Sand	Easement	1
123	21	PVC	18	0.07	6.00	801.57	7.9	Clay	Street	1
124	78	VCP	12	0.70	7.00	798.33	8.2	Clay	Alley	1
125	72	RC	8	1.00	7.00	796.06	7.9	Clay	Alley	1
126	71	RC	15	0.01	7.00	785.50	8.2	Clay	Highway	1
127	58	VCP	10	0.16	7.00	783.44	7.9	Clay	Highway	3
128	74	RC	78	0.06	10.00	783.29	7.9	Clay	Street	5
129	48	VCP	6	0.99	7.00	781.96	6.8	Sand	Alley	1
130	58	VCP	18	0.12	5.00	781.77	7.9	Clay	Street	1
131	74	RC	90	0.07	10.00	777.79	7.9	Clay	Alley	5
132	89	RC	6	0.60	7.00	776.67	7.9	Clay	Alley	4
133	52	VCP	6	0.02	8.00	774.65	8.2	Rock	Street	5
134	59	RC	18	0.22	7.00	773.71	6.8	Sand	Street	4
135	70	VCP	8	0.30	5.00	772.48	7.5	Clay	Street	1
136	54	VCP	8	0.20	8.00	770.75	6.8	Sand	Street	1
137	53	VCP	6	0.40	5.00	769.80	6.8	Sand	Alley	1
138	34	PVC	18	1.00	5.00	767.98	8.2	Clay	Street	1
139	66	RC	30	0.40	7.00	767.03	8.2	Loam	Highway	1
140	36	PVC	8	2.80	7.00	766.68	8.2	Clay	Street	1
141	81	RC	8	0.35	8.00	765.92	8.2	Clay	Street	4
142	56	RC	8	0.40	7.00	765.17	8.2	Clay	Street	1
143	71	RC	8	1.60	8.00	762.38	7.9	Clay	Street	4
144	65	RC	54	0.20	8.00	756.52	8.2	Loam	Easement	5
145	68	RC	33	0.10	6.00	753.54	7.5	Loam	Street	4
146	48	RC	30	0.34	11.00	752.64	8.2	Loam	Easement	1
147	39	PVC	15	0.10	10.00	752.50	7.9	Clay	Street	1
148	57	VCP	15	0.10	5.00	752.42	7.9	Clay	Street	1
149	13	PVC	24	0.75	10.00	751.95	8.2	Clay	Street	1
150	18	RC	66	0.03	7.00	751.57	7.9	Clay	Street	1
151	57	VCP	12	0.35	6.00	751.26	8.2	Clay	Street	1
152	39	RC	66	0.00	13.00	750.92	8.2	Loam	Highway	1
153	65	RC	30	0.20	5.00	749.54	8.2	Loam	Street	3
154	65	RC	10	2.20	8.00	745.24	8.2	Loam	Street	5
155	34	PVC	18	0.32	5.00	744.58	8.2	Clay	Easement	1
156	54	RC	10	0.40	10.00	741.78	8.2	Rock	Street	3
157	57	VCP	15	0.10	8.00	739.18	7.9	Clay	Alley	4
158	40	RC	66	0.09	13.00	739.03	8.2	Loam	Alley	1
159	30	PVC	10	1.00	7.00	736.11	7.9	Clay	Street	1
160	61	VCP	24	0.03	12.00	734.77	7.9	Clay	Street	1

161	64	RC	30	0.16	6.00	733.58	8.2	Loam	Street	5
162	67	RC	39	0.26	7.00	725.50	8.2	Loam	Alley	1
163	73	RC	24	0.35	10.00	719.70	8.2	Clay	Alley	5
164	54	RC	8	0.02	7.00	713.06	5.3	Sand	Alley	3
165	62	VCP	6	0.70	12.00	710.81	7.9	Clay	Street	2
166	65	RC	8	1.60	10.00	709.79	8.2	Clay	Street	3
167	65	VCP	8	0.40	12.00	709.70	7.9	Clay	Alley	3
168	39	RC	54	0.00	13.00	709.61	8.2	Loam	Street	1
169	49	VCP	15	0.50	12.00	709.30	7.9	Clay	Highway	1
170	60	VCP	8	0.20	5.00	707.55	7.9	Clay	Street	5
171	70	VCP	12	0.01	5.00	703.58	8.2	Clay	Alley	1
172	78	VCP	12	1.20	5.00	699.96	8.2	Clay	Easement	2
173	36	PVC	24	0.66	7.00	697.45	8.2	Clay	Street	3
174	50	VCP	6	0.90	5.00	695.64	6.8	Sand	Easement	3
175	55	VCP	10	0.08	6.00	695.50	5.8	Sand	Easement	3
176	43	PVC	8	0.01	7.00	695.45	8.2	Clay	Street	1
177	71	RC	8	0.40	7.00	694.68	7.9	Clay	Street	1
178	62	VCP	8	0.60	8.00	693.58	8.2	Clay	Street	1
179	63	VCP	15	0.20	7.00	693.48	7.9	Clay	Street	3
180	43	PVC	12	0.20	7.00	692.66	6.7	Sand	Street	5
181	70	RC	48	0.20	10.00	691.70	8.2	Loam	Street	5
182	60	RC	8	0.50	5.00	691.59	8.2	Clay	Highway	1
183	61	RC	6	1.20	5.00	690.54	8.2	Clay	Street	3
184	68	RC	33	0.10	6.00	690.26	7.5	Loam	Street	3
185	56	VCP	8	0.01	5.00	686.98	6.8	Sand	Street	3
186	79	VCP	6	1.00	7.00	686.14	8.2	Clay	Street	1
187	34	PVC	8	1.16	7.00	684.95	7.9	Clay	Street	1
188	66	VCP	8	0.32	7.00	684.23	7.9	Clay	Street	5
189	27	RC	42	0.10	6.00	681.27	7.9	Clay	Alley	5
190	37	PVC	12	0.50	15.00	679.79	7.9	Clay	Street	1
191	54	VCP	8	0.02	5.00	678.01	6.8	Sand	Street	5
192	58	RC	33	0.01	6.00	677.08	8.2	Loam	Street	1
193	65	RC	27	0.48	7.00	675.57	8.2	Loam	Alley	3
194	39	PVC	8	0.40	5.00	673.57	6.8	Sand	Street	1
195	45	RC	48	0.04	10.00	671.94	6.7	Sand	Easement	3
196	38	PVC	8	0.02	5.00	668.76	8.2	Clay	Street	1
197	61	VCP	18	0.26	7.00	667.81	6.8	Sand	Street	1
198	63	RC	8	0.80	7.00	667.66	7.9	Clay	Street	1
199	59	VCP	10	0.10	5.00	666.78	6.8	Sand	Alley	1
200	18	PVC	8	0.02	5.00	665.99	8.2	Clay	Easement	1
201	70	VCP	8	0.65	10.00	665.93	8.2	Clay	Street	5

202	44	RC	60	0.05	5.00	665.52	7.9	Clay	Highway	5
203	51	VCP	18	0.15	15.00	664.08	7.9	Clay	Easement	1
204	45	PVC	6	4.00	5.00	661.10	8.2	Clay	Street	1
205	15	PVC	10	0.30	10.00	660.26	8.2	Rock	Highway	1
206	26	PVC	8	0.75	8.00	658.14	7.9	Clay	Street	1
207	18	RC	66	0.03	7.00	657.26	7.9	Clay	Street	1
208	66	VCP	8	0.32	7.00	656.69	7.9	Clay	Street	2
209	79	RC	8	0.40	8.00	655.52	7.9	Clay	Street	1
210	65	RC	30	0.20	6.00	655.38	8.2	Loam	Street	3
211	63	RC	10	1.10	8.00	652.15	8.2	Clay	Street	1
212	42	PVC	15	0.38	10.00	651.20	7.5	Clay	Street	1
213	53	VCP	8	0.01	7.00	651.07	6.8	Sand	Alley	5
214	30	PVC	8	0.80	7.00	650.59	8.2	Clay	Street	1
215	36	PVC	8	2.20	6.00	649.95	7.9	Clay	Street	1
216	18	RC	66	0.03	5.00	649.60	7.9	Clay	Highway	1
217	73	RC	60	0.10	6.00	649.08	7.9	Clay	Alley	3
218	27	PVC	8	0.01	6.00	648.61	8.2	Clay	Street	1
219	54	VCP	8	0.30	8.00	645.89	6.8	Sand	Street	5
220	66	VCP	15	0.40	7.00	642.71	7.9	Clay	Street	1
221	70	VCP	18	0.29	7.00	639.80	7.9	Clay	Street	5
222	27	PVC	24	0.30	10.00	638.90	8.2	Loam	Street	1
223	38	PVC	10	0.33	5.00	638.31	6.8	Sand	Street	4
224	27	PVC	8	0.76	7.00	637.30	8.2	Clay	Alley	5
225	49	VCP	12	0.12	6.00	637.24	5.8	Sand	Street	1
226	21	PVC	15	0.68	7.00	635.53	8.2	Loam	Street	1
227	39	PVC	6	0.65	7.00	633.20	6.8	Sand	Highway	1
228	66	VCP	12	0.24	7.00	632.89	7.9	Clay	Street	1
229	8	PVC	12	0.20	5.00	632.75	7.9	Clay	Street	1
230	78	RC	8	0.40	8.00	629.51	8.2	Clay	Street	5
231	62	RC	6	0.01	8.00	624.96	8.2	Loam	Street	1
232	62	VCP	8	0.01	5.00	624.89	6.8	Sand	Street	5
233	29	PVC	18	0.01	8.00	624.73	7.9	Clay	Street	1
234	64	RC	12	1.00	8.00	623.66	8.2	Clay	Alley	1
235	65	RC	8	0.40	10.00	622.54	8.2	Rock	Street	5
236	66	VCP	15	1.20	7.00	622.35	7.9	Clay	Street	1
237	31	PVC	12	0.30	5.00	621.14	7.9	Clay	Alley	1
238	74	RC	90	0.05	10.00	620.09	7.9	Clay	Street	5
239	66	RC	27	0.30	6.00	619.34	7.9	Clay	Street	1
240	80	RC	10	0.02	10.00	618.42	8.2	Clay	Street	3
241	59	RC	12	0.90	13.00	617.19	8.2	Rock	Alley	1
242	64	VCP	8	0.12	8.00	617.01	7.9	Clay	Street	5

243	50	RC	60	0.16	6.00	616.85	7.9	Clay	Street	1
244	67	RC	10	1.00	10.00	616.47	8.2	Clay	Street	1
245	61	VCP	8	0.02	7.00	615.41	8.2	Clay	Alley	5
246	24	PVC	8	1.00	7.00	613.81	8.2	Clay	Street	1
247	46	VCP	6	0.60	5.00	613.40	7.9	Clay	Street	1
248	61	RC	8	0.03	10.00	612.48	8.2	Loam	Street	1
249	47	VCP	15	0.78	8.00	612.37	8.2	Loam	Street	2
250	62	RC	8	0.36	5.00	611.44	7.9	Clay	Highway	3
251	61	VCP	6	0.40	8.00	610.54	8.2	Clay	Street	1
252	49	VCP	10	1.10	10.00	609.09	8.2	Rock	Street	1
253	61	RC	15	0.02	13.00	607.95	8.2	Rock	Highway	1
254	67	VCP	24	0.07	7.00	607.54	7.9	Clay	Street	4
255	72	RC	8	1.00	7.00	603.61	8.2	Clay	Street	1
256	33	PVC	8	0.60	5.00	602.72	6.5	Sand	Alley	5
257	73	RC	60	0.09	10.00	602.52	7.9	Clay	Street	3
258	41	PVC	12	0.01	5.00	600.51	8.2	Loam	Easement	1
259	57	RC	18	0.01	7.00	599.68	8.2	Loam	Highway	3
260	47	VCP	6	0.02	8.00	599.11	8.2	Rock	Street	2
261	12	PVC	18	0.12	7.00	599.09	7.9	Clay	Street	1
262	72	RC	8	0.50	7.00	598.26	8.2	Clay	Street	1
263	61	VCP	18	0.20	7.00	597.65	6.8	Sand	Highway	1
264	67	VCP	27	0.50	7.00	596.99	7.9	Clay	Street	1
265	37	PVC	6	0.50	5.00	595.06	7.9	Clay	Street	1
266	27	PVC	8	1.00	10.00	594.46	7.9	Clay	Highway	1
267	65	RC	10	0.25	5.00	594.25	8.2	Loam	Street	1
268	16	PVC	24	0.40	7.00	594.21	7.9	Clay	Street	1
269	10	PVC	12	0.50	5.00	594.16	8.2	Clay	Alley	1
270	69	RC	10	0.20	7.00	593.98	7.9	Clay	Street	3
271	68	VCP	10	0.20	8.00	593.76	7.5	Clay	Street	4
272	23	PVC	8	0.45	5.00	592.54	8.2	Clay	Easement	1
273	67	RC	33	0.24	16.00	591.69	8.2	Loam	Highway	3
274	27	PVC	12	2.17	10.00	590.95	8.2	Loam	Easement	1
275	63	VCP	8	0.50	10.00	589.71	7.9	Clay	Street	1
276	67	VCP	8	0.00	10.00	589.56	8.2	Rock	Street	5
277	40	PVC	8	1.60	5.00	588.21	8.2	Clay	Street	1
278	11	PVC	18	0.11	5.00	588.20	8.2	Loam	Street	1
279	67	RC	24	0.82	10.00	587.89	8.2	Loam	Street	3
280	37	PVC	8	0.60	7.00	586.93	8.2	Clay	Alley	1
281	64	RC	24	0.12	6.00	586.34	7.5	Clay	Street	3
282	77	RC	8	0.01	8.00	584.49	8.2	Rock	Street	5
283	29	PVC	8	1.36	8.00	584.43	8.2	Clay	Alley	1

284	24	PVC	12	0.20	8.00	583.69	7.9	Clay	Street	1
285	34	PVC	8	0.72	10.00	583.22	7.9	Clay	Highway	1
286	71	RC	8	0.02	5.00	581.43	8.2	Clay	Street	4
287	59	RC	8	0.03	8.00	581.34	8.2	Rock	Street	5
288	14	PVC	8	0.61	15.00	581.22	7.9	Clay	Alley	1
289	55	RC	6	0.80	7.00	580.97	6.8	Sand	Street	3
290	66	RC	33	0.10	6.00	580.41	8.2	Clay	Street	1
291	36	PVC	10	0.60	6.00	580.11	7.9	Clay	Street	1
292	36	PVC	15	0.10	10.00	579.79	7.5	Loam	Street	1
293	12	PVC	12	0.15	7.00	578.88	7.9	Clay	Street	1
294	22	PVC	12	0.40	7.00	577.62	7.5	Loam	Street	1
295	68	VCP	8	0.20	5.00	576.69	7.5	Clay	Street	1
296	50	VCP	8	0.01	5.00	576.66	6.7	Sand	Street	1
297	77	RC	18	0.70	10.00	576.44	8.2	Loam	Easement	1
298	59	VCP	10	0.10	6.00	575.35	7.5	Clay	Alley	1
299	43	RC	54	0.07	5.00	572.12	8.2	Loam	Alley	5
300	30	PVC	24	0.36	7.00	569.47	7.5	Clay	Street	1
301	90	RC	42	0.10	7.00	569.08	6.7	Sand	Street	1
302	34	PVC	15	0.16	10.00	568.70	6.85	Loam	Street	1
303	91	RC	33	0.68	5.00	568.68	8.2	Loam	Street	3
304	84	RC	36	0.42	7.00	567.80	7.9	Clay	Alley	3
305	38	PVC	6	1.60	7.00	564.72	6.8	Sand	Street	1
306	27	PVC	10	0.36	16.00	564.62	7.9	Clay	Street	1
307	72	RC	8	0.03	5.00	563.93	8.2	Clay	Street	5
308	25	PVC	8	1.65	5.00	563.77	8.2	Clay	Street	1
309	52	VCP	15	0.10	6.00	563.68	5.8	Sand	Street	1
310	77	RC	6	0.09	7.00	563.34	8.2	Rock	Street	5
311	21	PVC	12	0.47	5.00	562.40	7.9	Clay	Alley	1
312	61	VCP	24	0.03	11.00	561.36	7.9	Clay	Street	1
313	55	VCP	8	0.60	10.00	560.86	6.8	Sand	Street	5
314	36	PVC	12	0.20	8.00	559.28	7.9	Clay	Alley	1
315	74	RC	72	0.08	10.00	558.93	7.9	Clay	Street	5
316	78	RC	18	0.36	7.00	557.48	8.2	Loam	Street	5
317	43	PVC	6	0.02	5.00	557.43	8.2	Clay	Highway	1
318	75	RC	8	1.60	7.00	557.20	8.2	Clay	Easement	1
319	35	PVC	10	0.30	15.00	557.03	7.9	Clay	Highway	1
320	66	RC	8	0.40	8.00	556.72	7.9	Clay	Alley	3
321	65	RC	54	0.20	10.00	555.85	8.2	Loam	Easement	3
322	50	VCP	12	1.20	7.00	555.51	5.3	Sand	Street	1
323	70	VCP	10	0.20	15.00	554.27	7.5	Clay	Street	5
324	62	VCP	10	0.20	7.00	554.05	6.8	Sand	Alley	2

325	21	PVC	12	0.47	5.00	553.82	7.9	Clay	Highway	1
326	89	VCP	6	1.00	5.00	552.17	8.2	Clay	Street	4
327	66	VCP	8	0.02	5.00	549.57	8.2	Rock	Street	1
328	20	PVC	18	0.50	5.00	549.45	8.2	Loam	Street	1
329	15	PVC	20	0.20	10.00	548.95	7.9	Clay	Street	1
330	59	RC	15	1.06	10.00	548.55	8.2	Loam	Street	1
331	51	VCP	12	0.40	10.00	547.87	7.9	Clay	Alley	3
332	81	RC	8	0.33	7.00	546.95	7.9	Clay	Highway	5
333	24	PVC	8	0.42	6.00	546.61	7.5	Clay	Highway	1
334	48	VCP	6	2.20	10.00	546.56	8.2	Rock	Street	4
335	40	PVC	15	0.30	10.00	545.52	7.5	Loam	Street	1
336	24	PVC	8	0.40	7.00	544.73	8.2	Clay	Street	1
337	70	RC	33	0.01	7.00	544.65	8.2	Loam	Alley	5
338	51	VCP	8	0.00	5.00	542.81	8.2	Clay	Street	5
339	66	RC	24	0.42	7.00	542.17	8.2	Loam	Street	1
340	25	PVC	8	0.50	7.00	541.99	7.9	Clay	Street	4
341	41	PVC	10	0.40	15.00	541.76	7.9	Clay	Highway	1
342	22	PVC	8	2.20	7.00	540.73	8.2	Clay	Street	1
343	90	RC	42	0.10	7.00	540.48	6.7	Sand	Street	3
344	59	RC	6	0.06	7.00	539.73	8.2	Rock	Easement	2
345	11	PVC	42	0.21	7.00	537.98	8.2	Loam	Street	3
346	25	PVC	30	0.08	7.00	536.83	8.2	Loam	Street	1
347	50	VCP	8	0.30	11.00	535.96	6.7	Sand	Street	1
348	63	RC	8	0.02	7.00	535.91	8.2	Rock	Street	1
349	66	RC	24	0.60	7.00	535.00	8.2	Loam	Highway	3
350	51	RC	8	0.50	10.00	534.84	7.9	Clay	Street	1
351	35	PVC	8	0.80	10.00	534.67	6.8	Sand	Street	1
352	34	PVC	8	0.40	15.00	533.31	7.9	Clay	Street	1
353	35	PVC	8	0.50	7.00	532.46	8.2	Clay	Easement	1
354	48	VCP	12	2.40	7.00	532.37	6.7	Sand	Street	1
355	60	RC	6	1.20	8.00	531.80	8.2	Clay	Easement	3
356	66	RC	30	0.08	5.00	531.44	8.2	Loam	Street	1
357	18	PVC	8	0.40	8.00	530.72	8.2	Clay	Street	1
358	54	VCP	15	0.30	10.00	530.38	7.9	Clay	Street	1
359	36	PVC	24	0.76	5.00	530.35	8.2	Clay	Street	1
360	40	PVC	6	0.48	7.00	529.50	7.9	Clay	Street	1
361	25	PVC	8	1.20	5.00	529.08	7.9	Clay	Highway	1
362	59	VCP	12	0.14	13.00	526.98	7.9	Clay	Street	1
363	38	PVC	8	1.33	5.00	526.54	8.2	Clay	Street	1
364	71	VCP	8	0.20	8.00	526.02	7.5	Clay	Alley	1
365	24	PVC	8	0.46	7.00	525.14	7.9	Clay	Alley	1

366	16	PVC	8	0.33	7.00	524.91	8.2	Clay	Street	1
367	49	VCP	6	0.02	10.00	524.30	8.2	Clay	Street	1
368	29	PVC	8	0.02	5.00	523.00	7.9	Clay	Street	1
369	45	VCP	12	0.36	10.00	522.94	7.9	Clay	Street	1
370	26	VCP	30	0.20	14.06	522.59	5.4	Gravel	Alley	1
371	23	PVC	8	0.50	5.00	522.41	8.2	Clay	Street	1
372	10	PVC	8	0.84	7.00	522.39	7.9	Clay	Street	1
373	64	RC	36	0.10	7.00	521.94	8.2	Loam	Street	1
374	24	PVC	8	2.20	7.00	521.35	8.2	Clay	Street	1
375	54	VCP	15	0.30	10.00	520.96	7.9	Clay	Street	1
376	25	PVC	8	1.23	5.00	520.08	8.2	Clay	Highway	1
377	22	PVC	12	0.40	7.00	519.95	7.5	Loam	Street	1
378	47	VCP	15	1.80	10.00	519.32	8.2	Loam	Highway	1
379	16	PVC	12	0.20	7.00	518.74	7.9	Clay	Easement	5
380	68	RC	8	1.06	7.00	518.68	7.9	Clay	Highway	5
381	27	PVC	8	0.90	8.00	517.74	7.5	Clay	Alley	1
382	74	RC	36	0.20	10.00	517.18	6.7	Sand	Street	3
383	22	PVC	8	0.34	6.00	516.97	7.9	Clay	Street	1
384	45	VCP	15	0.01	7.00	516.47	7.5	Loam	Street	1
385	48	VCP	12	0.20	11.00	515.93	7.9	Clay	Street	3
386	37	PVC	6	0.80	12.00	515.79	7.5	Clay	Street	1
387	47	VCP	8	0.02	8.00	515.12	8.2	Rock	Street	5
388	51	VCP	8	3.00	7.00	514.20	6.5	Sand	Street	3
389	46	VCP	6	0.60	7.00	513.94	8.2	Clay	Alley	1
390	23	PVC	8	0.66	5.00	512.87	8.2	Loam	Street	1
391	61	RC	54	0.40	8.00	512.62	8.2	Loam	Highway	3
392	58	RC	21	3.60	10.00	512.07	7.9	Clay	Street	5
393	41	PVC	8	0.52	10.00	511.24	7.9	Clay	Street	1
394	90	RC	42	0.10	7.00	510.76	7.5	Clay	Street	3
395	89	VCP	8	0.30	10.00	509.07	8.2	Clay	Alley	5
396	75	RC	42	0.10	5.00	508.57	7.5	Clay	Street	5
397	17	PVC	8	1.40	5.00	507.55	8.2	Clay	Street	1
398	40	VCP	10	0.30	10.00	507.48	6.85	Loam	Street	1
399	63	RC	6	0.60	8.00	507.32	8.2	Loam	Street	1
400	17	PVC	8	1.40	15.00	506.69	7.9	Clay	Alley	1
401	67	VCP	27	0.50	7.00	506.57	7.9	Clay	Alley	1
402	73	RC	15	0.36	5.00	506.46	8.2	Clay	Street	1
403	22	PVC	12	0.40	7.00	506.22	7.9	Clay	Street	1
404	48	VCP	6	0.50	7.00	504.82	8.2	Loam	Highway	3
405	23	PVC	8	0.56	8.00	504.64	7.9	Clay	Street	1
406	18	PVC	8	2.30	8.00	504.53	8.2	Rock	Easement	1

407	61	VCP	8	0.30	10.00	503.57	7.9	Clay	Street	1
408	67	RC	27	0.58	7.00	503.25	8.2	Loam	Highway	3
409	23	PVC	8	1.00	5.00	503.01	7.9	Clay	Street	2
410	35	PVC	8	0.40	15.00	502.66	7.9	Clay	Street	1
411	15	PVC	8	0.40	6.00	501.27	7.9	Clay	Street	1
412	53	RC	39	0.47	11.00	501.18	8.2	Loam	Street	1
413	12	PVC	8	2.08	8.00	500.78	8.2	Clay	Street	1
414	73	RC	54	0.16	6.00	500.73	8.2	Loam	Street	5
415	11	PVC	8	0.35	8.00	500.45	8.2	Clay	Street	1
416	40	PVC	6	0.60	10.00	500.39	6.85	Loam	Street	1
417	70	VCP	8	0.02	7.00	500.36	8.2	Clay	Highway	5
418	8	PVC	12	0.20	7.00	500.24	6.8	Sand	Alley	1
419	18	PVC	8	0.00	19.54	500.02	6.5	Sand	Highway	3
420	12	PVC	8	1.66	8.00	499.88	8.2	Clay	Street	5
421	11	PVC	8	3.80	8.00	499.71	8.2	Clay	Highway	1
422	21	PVC	8	2.00	5.00	499.62	8.2	Clay	Street	1
423	14	PVC	8	2.00	5.00	499.32	7.9	Clay	Street	1
424	33	PVC	10	1.28	10.00	499.09	8.2	Clay	Street	1
425	21	PVC	8	0.28	7.00	498.80	7.5	Clay	Street	1
426	14	PVC	8	0.48	7.00	498.76	7.5	Clay	Alley	1
427	18	PVC	12	1.78	7.00	498.25	8.2	Clay	Street	1
428	27	PVC	8	0.45	5.00	497.98	8.2	Clay	Street	1
429	24	PVC	8	1.90	6.00	497.60	8.2	Clay	Street	1
430	13	PVC	8	1.50	7.00	497.33	8.2	Clay	Street	1
431	14	PVC	12	1.06	7.00	497.11	6.7	Sand	Street	1
432	47	VCP	8	1.30	10.00	496.91	7.9	Clay	Street	5
433	47	VCP	12	0.60	7.00	496.84	8.2	Loam	Street	1
434	18	PVC	8	1.50	5.00	496.83	7.9	Clay	Street	2
435	18	PVC	12	0.30	5.00	496.73	8.2	Clay	Street	1
436	60	VCP	21	0.10	7.00	495.92	7.9	Clay	Highway	1
437	21	PVC	10	0.50	7.00	495.81	7.5	Clay	Street	1
438	18	RC	66	0.03	8.00	495.76	7.9	Clay	Street	3
439	35	PVC	15	0.60	10.00	495.72	7.5	Clay	Street	1
440	66	VCP	8	3.50	7.00	495.71	8.2	Rock	Alley	1
441	10	PVC	8	0.50	7.00	495.70	8.2	Clay	Street	1
442	44	PVC	8	0.30	10.00	495.56	8.2	Loam	Street	1
443	48	VCP	6	4.03	8.00	495.54	8.2	Clay	Street	1
444	14	PVC	12	1.12	5.00	495.46	8.2	Clay	Street	1
445	17	PVC	8	1.80	10.00	495.00	8.2	Clay	Alley	1
446	21	PVC	8	1.40	5.00	493.73	7.9	Clay	Highway	1
447	35	PVC	8	1.20	5.00	493.69	6.8	Sand	Alley	1

448	59	RC	12	0.62	13.00	493.55	8.2	Rock	Street	3
449	67	VCP	8	0.30	7.00	492.53	7.5	Clay	Street	2
450	47	VCP	6	0.60	8.00	492.44	8.2	Clay	Alley	1
451	49	VCP	10	0.30	5.00	492.14	6.8	Sand	Easement	1
452	17	PVC	8	1.58	7.00	491.74	8.2	Clay	Street	1
453	55	RC	15	1.20	7.00	491.61	8.2	Loam	Street	3
454	11	PVC	8	3.00	6.00	491.35	7.9	Clay	Street	1
455	20	PVC	8	1.91	10.00	491.27	7.9	Clay	Street	1
456	35	PVC	8	0.80	10.00	491.22	8.2	Rock	Street	1
457	65	RC	24	0.50	7.00	491.16	8.2	Loam	Alley	1
458	63	VCP	15	0.12	5.00	491.07	6.8	Sand	Street	3
459	42	PVC	15	0.26	10.00	490.96	7	Loam	Street	1
460	78	RC	21	0.70	10.00	490.61	8.2	Loam	Easement	1
461	20	PVC	8	1.42	5.00	490.60	7.9	Clay	Alley	1
462	61	VCP	24	0.03	12.00	489.57	7.9	Clay	Street	1
463	56	VCP	8	0.25	7.00	488.22	7.5	Loam	Alley	4
464	62	RC	8	0.36	8.00	486.77	7.9	Clay	Alley	3
465	66	RC	10	0.36	5.00	486.47	8.2	Loam	Street	1
466	47	VCP	15	0.83	10.00	486.35	7.9	Clay	Street	1
467	29	PVC	8	0.02	10.00	486.22	7.9	Clay	Alley	1
468	11	PVC	8	1.06	7.00	486.18	6.8	Sand	Street	1
469	23	PVC	8	0.81	8.00	485.71	8.2	Clay	Street	1
470	22	PVC	12	0.56	7.00	485.49	7.9	Clay	Street	1
471	14	PVC	8	0.35	5.00	485.07	8.2	Clay	Street	1
472	21	PVC	8	0.72	7.00	484.74	8.2	Loam	Highway	1
473	65	RC	54	0.20	6.00	484.41	8.2	Loam	Street	3
474	54	RC	18	0.48	15.00	484.14	7.9	Clay	Street	1
475	45	VCP	10	0.18	6.00	482.81	5.5	Sand	Alley	1
476	36	PVC	6	2.40	8.00	482.53	6.8	Sand	Street	1
477	63	VCP	8	5.00	7.00	481.48	8.2	Clay	Street	1
478	64	VCP	8	0.30	15.00	481.15	7.5	Clay	Street	4
479	12	PVC	8	0.33	5.00	481.09	6.7	Sand	Street	1
480	29	PVC	48	0.20	10.00	480.49	8.2	Loam	Street	1
481	43	PVC	12	0.24	15.00	480.16	6.7	Sand	Street	1
482	9	PVC	12	0.20	8.00	479.98	6.7	Sand	Street	1
483	18	PVC	8	1.69	7.00	479.96	8.2	Clay	Street	5
484	54	VCP	18	0.48	15.00	479.87	7.9	Clay	Street	1
485	36	PVC	15	0.40	10.00	478.68	8.2	Loam	Street	1
486	17	PVC	10	0.26	5.00	478.65	7.5	Clay	Street	1
487	40	PVC	10	0.01	6.00	476.89	5.8	Sand	Easement	5
488	22	PVC	8	0.60	5.00	476.76	7.9	Clay	Alley	1

489	72	RC	8	0.02	7.00	476.64	7.9	Clay	Street	1
490	70	VCP	12	0.01	7.00	476.43	8.2	Clay	Street	5
491	21	PVC	8	1.12	7.00	476.27	7.9	Clay	Street	1
492	44	PVC	12	0.08	6.00	475.21	5.8	Sand	Street	1
493	61	VCP	6	0.71	5.00	474.62	6.8	Sand	Street	5
494	67	VCP	8	0.30	10.00	474.52	7.9	Clay	Highway	5
495	12	PVC	8	0.33	5.00	474.40	6.7	Sand	Street	1
496	25	PVC	8	1.51	8.00	474.02	8.2	Loam	Street	1
497	75	RC	8	1.20	7.00	473.27	7.9	Clay	Street	1
498	12	PVC	8	0.75	6.00	472.75	7.9	Clay	Street	1
499	66	RC	30	0.16	7.00	472.61	7.9	Clay	Alley	4
500	18	PVC	8	2.24	8.00	472.54	8.2	Clay	Highway	1
501	53	VCP	10	1.00	10.00	472.49	7	Loam	Easement	1
502	71	RC	12	0.60	10.00	472.14	6.7	Sand	Alley	1
503	13	PVC	15	0.15	8.00	471.09	7.9	Clay	Alley	1
504	12	PVC	18	0.12	10.00	470.92	7.9	Clay	Street	1
505	20	PVC	8	1.39	4.00	470.81	8.2	Clay	Street	1
506	84	RC	36	0.42	7.00	470.72	7.9	Clay	Easement	3
507	18	PVC	8	0.40	7.00	470.40	7.9	Clay	Street	1
508	57	VCP	10	0.50	8.00	469.73	8.2	Loam	Street	1
509	59	RC	10	0.40	7.00	469.38	6.8	Sand	Alley	2
510	58	RC	6	0.60	10.00	468.93	7.9	Clay	Street	1
511	26	PVC	8	0.01	6.00	467.72	8.2	Clay	Street	1
512	18	PVC	10	0.25	5.00	467.36	7.9	Clay	Street	1
513	67	RC	30	0.64	8.00	467.06	8.2	Loam	Easement	3
514	30	PVC	8	0.50	6.00	466.75	7.9	Clay	Street	1
515	43	PVC	8	0.34	8.00	466.12	7.9	Clay	Street	1
516	17	PVC	8	0.75	15.00	465.77	7.9	Clay	Highway	1
517	73	VCP	8	0.40	7.00	465.64	6.7	Sand	Street	1
518	56	VCP	10	1.70	7.00	465.60	8.2	Loam	Street	1
519	64	RC	24	0.12	10.00	465.47	6.7	Sand	Alley	1
520	54	VCP	21	0.34	15.00	464.55	7.9	Clay	Street	1
521	51	VCP	8	1.00	7.00	464.39	7.9	Clay	Alley	1
522	24	PVC	8	0.60	3.00	463.68	8.2	Clay	Street	1
523	15	PVC	8	0.40	5.00	463.09	6.8	Sand	Street	1
524	68	VCP	8	0.40	7.00	462.26	8.2	Loam	Street	2
525	64	VCP	8	0.28	8.00	462.17	8.2	Loam	Street	5
526	72	VCP	8	0.03	5.00	461.81	8.2	Loam	Street	5
527	54	VCP	8	0.20	5.00	461.68	6.8	Sand	Street	1
528	62	VCP	12	0.50	12.00	461.06	7.9	Clay	Street	3
529	56	VCP	8	0.20	5.00	461.05	6.8	Sand	Highway	1

530	40	PVC	6	1.20	7.00	460.90	8.2	Loam	Street	2
531	50	VCP	15	0.01	7.00	459.81	6.8	Sand	Street	1
532	16	PVC	8	1.33	5.00	458.70	7.9	Clay	Street	1
533	43	RC	54	0.06	8.00	458.60	7.9	Clay	Easement	5
534	67	VCP	24	0.03	8.00	458.33	7.9	Clay	Street	3
535	21	PVC	8	0.50	5.00	457.96	7.9	Clay	Easement	1
536	24	PVC	8	0.73	6.00	457.95	7.5	Clay	Street	1
537	23	PVC	8	1.10	6.00	457.64	7.9	Clay	Street	1
538	72	RC	8	0.35	7.00	457.58	7.9	Clay	Street	5
539	70	VCP	8	1.76	7.00	457.42	8.2	Clay	Street	4
540	25	PVC	8	0.60	7.00	457.11	7.5	Clay	Alley	1
541	79	RC	30	0.06	5.00	456.84	7.9	Clay	Street	1
542	15	PVC	8	0.75	10.00	456.62	7	Loam	Street	1
543	33	PVC	8	0.74	6.00	456.34	7.9	Clay	Street	1
544	13	PVC	15	0.60	7.00	456.20	7.9	Clay	Street	1
545	23	PVC	8	1.12	8.00	455.93	7.9	Clay	Street	1
546	61	RC	6	0.60	10.00	455.76	8.2	Clay	Street	1
547	35	PVC	18	0.50	15.00	454.85	7.5	Clay	Easement	1
548	63	RC	27	0.32	6.00	452.92	7.9	Clay	Street	1
549	15	PVC	8	0.40	7.00	452.52	7.9	Clay	Street	1
550	46	VCP	6	2.40	8.00	451.90	8.2	Loam	Easement	1
551	55	VCP	10	0.60	7.00	451.50	7.9	Clay	Street	1
552	67	RC	30	0.20	15.00	451.14	7.9	Clay	Street	5
553	15	PVC	18	0.30	5.00	450.84	8.2	Loam	Easement	1
554	41	PVC	6	2.20	8.00	450.82	8.2	Rock	Street	1
555	59	VCP	10	0.90	5.00	450.58	6.8	Sand	Street	3
556	23	PVC	8	0.35	5.00	450.41	7.9	Clay	Street	1
557	67	VCP	8	0.29	11.59	450.00	7.4	Sand	Street	3
558	10	PVC	8	0.50	8.00	449.96	8.2	Loam	Street	1
559	14	PVC	8	0.34	7.00	449.78	7.9	Clay	Alley	2
560	48	VCP	8	0.01	8.00	449.69	7.9	Clay	Street	5
561	12	PVC	8	0.75	5.00	449.14	7.9	Clay	Street	3
562	13	PVC	8	4.47	10.00	449.05	8.2	Clay	Street	1
563	63	VCP	15	0.18	10.00	448.70	6.8	Sand	Street	1
564	83	RC	36	0.26	7.00	448.40	7.9	Clay	Easement	1
565	78	RC	21	0.70	10.00	448.30	8.2	Loam	Street	1
566	8	PVC	8	0.33	6.00	447.34	8.2	Clay	Street	1
567	62	VCP	12	0.20	12.00	447.05	7.9	Clay	Alley	3
568	12	PVC	8	1.90	7.00	446.75	8.2	Clay	Street	1
569	23	PVC	8	0.84	5.00	446.53	8.2	Clay	Street	2
570	11	PVC	8	0.50	7.00	445.67	8.2	Clay	Street	1

571	16	PVC	8	0.40	8.00	445.31	6.8	Sand	Highway	1
572	65	VCP	8	0.36	10.00	445.29	8.2	Clay	Alley	3
573	60	VCP	12	0.20	7.00	444.74	7.9	Clay	Highway	2
574	27	PVC	24	1.15	10.00	443.92	8.2	Loam	Street	1
575	111	VCP	8	1.14	7.00	443.70	5.3	Sand	Street	4
576	48	VCP	12	0.30	8.00	443.01	6.7	Sand	Street	1
577	23	PVC	8	2.80	5.00	442.43	8.2	Clay	Street	1
578	14	PVC	8	2.00	7.00	441.48	6.8	Sand	Street	1
579	21	PVC	8	0.40	8.00	440.42	7.5	Clay	Alley	1
580	48	VCP	15	0.84	12.00	440.33	7.9	Clay	Street	1
581	70	RC	10	0.80	7.00	439.70	8.2	Loam	Street	3
582	62	VCP	12	0.20	12.00	439.62	7.9	Clay	Street	1
583	21	PVC	8	0.40	6.00	439.25	7.9	Clay	Street	1
584	25	PVC	12	1.00	8.00	438.84	8.2	Clay	Street	1
585	52	RC	6	0.90	8.00	438.41	6.8	Sand	Alley	1
586	59	VCP	8	0.25	5.00	438.05	7.9	Clay	Street	1
587	75	RC	8	0.60	7.00	437.56	7.9	Clay	Street	5
588	36	PVC	10	1.50	10.00	437.49	8.2	Loam	Street	1
589	31	PVC	18	0.32	7.00	437.44	7.9	Clay	Street	1
590	23	PVC	8	0.37	5.00	436.95	7.9	Clay	Street	1
591	56	VCP	8	0.25	6.00	436.65	7.5	Clay	Street	1
592	67	RC	8	0.03	7.00	436.14	7.9	Clay	Street	3
593	71	VCP	8	0.30	11.32	435.58	5.7	Sand	Street	3
594	30	PVC	8	0.30	8.00	435.45	7.9	Clay	Highway	1
595	17	PVC	8	3.20	8.00	435.43	8.2	Clay	Alley	1
596	32	PVC	8	0.80	5.00	435.26	7.9	Clay	Street	1
597	61	RC	18	0.50	10.00	434.91	8.2	Clay	Street	3
598	12	PVC	8	1.08	9.00	434.89	8.2	Rock	Street	1
599	21	PVC	8	1.42	7.00	434.28	7.9	Clay	Street	1
600	22	PVC	8	3.20	5.00	434.10	8.2	Clay	Street	1
601	46	PVC	8	0.52	7.00	433.16	7.9	Clay	Street	1
602	62	RC	8	0.90	5.00	432.81	8.2	Rock	Street	3
603	11	PVC	8	2.99	7.00	432.25	8.2	Clay	Street	1
604	95	VCP	12	0.30	5.00	431.76	7.9	Clay	Street	1
605	57	VCP	12	0.02	7.00	431.61	8.2	Clay	Street	1
606	62	VCP	6	0.60	8.00	431.50	7.9	Clay	Alley	1
607	29	PVC	8	1.73	6.00	430.64	8.2	Clay	Street	1
608	24	PVC	8	1.30	8.00	430.19	8.2	Clay	Street	1
609	18	PVC	8	1.00	5.00	430.00	8.2	Clay	Street	1
610	91	RC	33	0.72	7.00	429.80	8.2	Loam	Easement	3
611	16	PVC	8	1.26	5.00	429.57	8.2	Clay	Street	1

612	70	VCP	8	1.80	7.00	429.47	7.9	Clay	Easement	4
613	56	RC	8	1.30	7.00	429.40	8.2	Clay	Street	1
614	39	RC	18	0.10	6.00	429.23	8.2	Clay	Street	3
615	95	VCP	12	0.30	5.00	429.09	7.9	Clay	Street	1
616	62	VCP	12	0.50	8.00	428.70	7.9	Clay	Alley	2
617	95	VCP	12	0.30	5.00	428.56	7.9	Clay	Easement	1
618	20	PVC	8	0.33	7.00	428.40	7.9	Clay	Street	1
619	26	PVC	8	1.50	8.00	428.29	8.2	Clay	Alley	1
620	73	RC	72	0.06	10.00	428.14	7.9	Clay	Street	3
621	17	PVC	10	0.88	5.00	427.69	6.8	Sand	Street	1
622	69	RC	30	0.30	6.00	427.57	7.9	Clay	Street	5
623	90	RC	42	0.40	5.00	427.53	6.7	Sand	Street	5
624	19	PVC	10	0.26	7.00	427.26	7.9	Clay	Street	1
625	29	PVC	16	0.40	16.00	426.91	7.9	Clay	Alley	1
626	22	PVC	12	0.40	8.00	426.69	7.9	Clay	Easement	1
627	12	PVC	8	0.03	7.00	425.70	8.2	Clay	Highway	1
628	58	VCP	8	0.69	0.64	425.40	5.7	Sand	Street	3
629	63	RC	10	0.40	10.00	425.35	8.2	Loam	Street	2
630	22	PVC	8	1.06	7.00	425.31	8.2	Clay	Street	1
631	17	PVC	8	1.06	5.00	425.20	8.2	Clay	Street	1
632	73	RC	10	0.93	10.00	424.31	7.9	Clay	Alley	1
633	24	PVC	8	4.35	6.00	424.17	8.2	Clay	Easement	1
634	20	PVC	8	0.99	5.00	423.85	7.9	Clay	Street	1
635	53	VCP	10	0.30	15.00	423.77	7.3	Clay	Street	3
636	37	PVC	12	0.50	15.00	423.64	7.9	Clay	Street	1
637	34	PVC	8	0.40	5.00	423.58	6.8	Sand	Street	1
638	36	PVC	8	3.98	5.00	423.54	6.8	Sand	Street	1
639	40	PVC	10	0.10	6.00	423.28	5.8	Sand	Street	1
640	18	PVC	8	0.40	7.00	423.25	8.2	Clay	Street	1
641	46	VCP	8	0.50	7.00	423.25	7.5	Clay	Street	1
642	58	VCP	10	0.30	7.00	423.01	7.9	Clay	Highway	2
643	64	VCP	8	0.28	7.00	422.98	8.2	Loam	Street	2
644	23	PVC	8	0.50	7.00	422.70	8.2	Loam	Street	1
645	33	PVC	8	0.80	7.00	422.49	6.5	Sand	Alley	1
646	14	PVC	8	0.36	8.00	422.43	7.5	Loam	Street	1
647	35	PVC	15	2.20	11.00	422.30	7.5	Clay	Easement	1
648	14	PVC	8	0.33	8.00	421.71	7.5	Clay	Street	1
649	75	RC	27	1.00	5.00	421.07	8.2	Clay	Highway	5
650	50	VCP	10	0.10	15.00	421.04	6.7	Sand	Street	1
651	77	RC	18	0.55	8.00	420.43	8.2	Loam	Street	1
652	45	VCP	8	0.20	6.00	420.35	6.5	Sand	Street	1

653	35	PVC	8	0.50	5.00	420.03	8.2	Clay	Highway	1
654	18	PVC	8	1.00	5.00	419.48	8.2	Clay	Street	1
655	26	PVC	8	0.02	6.00	419.40	8.2	Clay	Alley	1
656	37	VCP	15	0.14	15.70	418.92	4.8	Sand	Street	3
657	60	VCP	6	1.00	5.00	418.58	6.8	Sand	Street	1
658	14	PVC	8	0.40	5.00	418.27	7.9	Clay	Street	1
659	38	PVC	12	3.82	7.00	418.21	5.3	Sand	Highway	1
660	38	PVC	8	0.40	10.00	416.77	8.2	Rock	Street	1
661	17	PVC	8	0.80	8.00	416.27	8.2	Clay	Easement	1
662	73	RC	60	0.06	10.00	415.93	7.9	Clay	Street	3
663	57	VCP	6	0.50	7.00	415.48	8.2	Clay	Highway	1
664	23	PVC	8	0.86	7.00	415.34	6.8	Sand	Street	1
665	62	VCP	12	0.50	8.00	414.95	7.9	Clay	Street	3
666	58	VCP	8	0.32	10.77	414.79	5.4	Sand	Street	4
667	77	VCP	6	2.20	7.00	414.62	8.2	Clay	Street	5
668	39	PVC	15	0.10	6.00	414.50	7.9	Clay	Street	1
669	8	PVC	12	0.65	5.00	414.03	8.2	Clay	Street	1
670	14	PVC	8	2.60	7.00	414.02	8.2	Clay	Alley	1
671	18	PVC	8	0.72	7.00	413.66	8.2	Clay	Street	1
672	20	PVC	8	0.50	6.00	412.83	7.9	Clay	Street	1
673	62	RC	6	0.60	12.00	412.76	7.9	Clay	Street	1
674	74	RC	8	0.40	10.00	412.63	8.2	Clay	Street	4
675	72	VCP	8	0.35	8.00	412.29	7.5	Clay	Highway	1
676	41	PVC	12	1.00	10.00	411.71	8.2	Loam	Street	1
677	101	VCP	8	0.30	7.00	411.51	8.2	Clay	Street	5
678	103	VCP	8	0.30	5.00	410.96	8.2	Clay	Street	3
679	67	RC	10	0.80	8.00	410.18	8.2	Loam	Street	3
680	45	VCP	8	0.40	8.58	410.00	4.8	Sand	Alley	3
681	53	RC	8	0.40	7.00	408.49	8.2	Rock	Street	1
682	38	PVC	8	1.34	10.00	408.30	8.2	Loam	Street	3
683	62	RC	30	0.60	7.00	407.91	8.2	Loam	Street	3
684	42	PVC	8	0.30	10.00	407.72	7	Loam	Street	1
685	50	VCP	8	0.40	8.58	407.35	5.4	Sand	Street	2
686	41	PVC	10	0.98	5.00	407.30	8.2	Loam	Street	1
687	23	PVC	8	0.35	5.00	407.09	7.9	Clay	Street	1
688	53	VCP	8	0.40	5.00	407.07	7.9	Clay	Street	1
689	10	PVC	12	0.20	6.00	406.98	5.8	Sand	Street	1
690	44	PVC	12	0.08	6.00	406.96	5.8	Sand	Street	1
691	38	PVC	8	0.50	12.00	406.80	7.5	Clay	Street	1
692	8	PVC	12	0.45	5.00	406.43	8.2	Clay	Street	1
693	77	RC	21	0.85	7.00	406.09	8.2	Loam	Street	1

694	16	PVC	8	0.44	7.00	406.06	8.2	Clay	Easement	1
695	69	VCP	8	0.22	13.51	405.89	4.8	Sand	Street	3
696	9	PVC	8	1.20	5.00	405.86	6.8	Sand	Street	1
697	36	PVC	15	0.60	10.00	405.80	7.5	Loam	Street	1
698	24	PVC	8	1.60	7.00	405.72	8.2	Clay	Street	1
699	51	RC	6	0.60	5.00	405.37	6.8	Sand	Highway	1
700	13	PVC	8	1.74	8.00	404.43	7.9	Clay	Street	1
701	29	PVC	8	0.80	8.00	403.97	7.9	Clay	Street	1
702	54	RC	21	0.05	7.00	403.69	8.2	Clay	Alley	1
703	93	VCP	8	0.60	7.00	403.40	6.7	Sand	Street	5
704	13	PVC	8	1.00	5.00	402.95	8.2	Clay	Street	1
705	18	PVC	8	0.03	8.00	402.80	8.2	Clay	Street	1
706	18	PVC	8	0.33	6.00	402.79	8.2	Clay	Street	1
707	101	VCP	6	3.70	7.00	402.74	8.2	Clay	Street	5
708	17	PVC	8	1.10	7.00	402.69	8.2	Clay	Street	1
709	43	RC	27	0.32	7.00	402.65	8.2	Clay	Street	1
710	55	RC	8	0.84	7.00	402.38	6.8	Sand	Street	1
711	46	VCP	8	0.34	10.22	402.17	5.4	Sand	Street	3
712	24	PVC	8	1.50	5.00	402.15	8.2	Clay	Highway	1
713	38	PVC	8	1.32	12.00	401.85	7.5	Clay	Alley	1
714	13	PVC	12	2.50	7.00	401.83	8.2	Loam	Street	1
715	13	PVC	18	0.15	7.00	401.74	6.8	Sand	Highway	1
716	58	RC	15	1.88	10.00	401.69	8.2	Loam	Street	1
717	59	RC	6	1.00	10.00	401.61	8.2	Clay	Street	3
718	15	PVC	12	0.40	5.00	401.58	8.2	Rock	Street	1
719	42	VCP	8	0.27	12.14	401.49	6.5	Sand	Street	3
720	22	PVC	8	0.36	6.00	401.46	7.9	Clay	Alley	1
721	51	VCP	12	0.30	10.00	401.40	6.7	Sand	Street	1
722	57	VCP	8	0.41	8.31	401.39	5.4	Sand	Street	3
723	57	VCP	8	0.62	2.55	401.30	6.9	Sand	Easement	3
724	57	VCP	8	0.46	6.94	401.00	4.1	Sand	Street	3
725	57	VCP	8	0.39	8.85	400.99	7.4	Sand	Highway	3
726	60	VCP	8	0.38	7.00	400.94	6.8	Sand	Alley	1
727	46	VCP	8	0.46	6.94	400.93	5.4	Sand	Street	3
728	63	VCP	15	0.20	7.00	400.88	7.9	Clay	Street	1
729	50	VCP	8	0.60	15.00	400.76	7.9	Clay	Street	1
730	67	VCP	8	0.40	7.00	400.69	8.2	Loam	Street	1
731	58	VCP	10	0.31	11.05	400.59	5.7	Sand	Street	3
732	58	VCP	10	0.23	13.24	400.49	5.7	Sand	Street	3
733	58	VCP	8	0.31	11.05	400.49	5.4	Sand	Street	3
734	62	VCP	8	0.36	9.00	400.42	7.9	Clay	Street	2

735	69	VCP	8	0.34	10.22	400.10	4.8	Sand	Highway	3
736	21	PVC	8	2.50	7.00	400.06	8.2	Loam	Street	1
737	71	VCP	8	0.30	11.32	400.00	5.7	Sand	Alley	3
738	52	VCP	8	0.47	6.66	400.00	5.7	Sand	Highway	3
739	69	VCP	8	0.30	11.32	399.99	7.9	Sand	Street	3
740	69	VCP	8	0.36	9.68	399.99	5.4	Sand	Street	3
741	49	VCP	8	0.40	8.58	399.99	5.4	Sand	Alley	1
742	58	VCP	8	0.47	6.66	399.99	5.4	Sand	Street	3
743	11	PVC	8	1.40	8.00	399.98	8.2	Clay	Street	1
744	69	VCP	8	0.40	8.58	399.97	4.8	Sand	Street	3
745	43	PVC	8	0.40	10.00	399.88	7.9	Clay	Street	3
746	46	VCP	8	0.34	10.22	399.86	5.8	Sand	Alley	3
747	17	PVC	8	4.20	5.00	399.71	8.2	Loam	Street	1
748	69	VCP	8	0.40	8.58	399.57	5.4	Sand	Street	3
749	57	VCP	8	0.47	6.66	399.49	5.4	Sand	Alley	3
750	24	PVC	8	0.40	10.00	399.47	7.9	Clay	Street	1
751	16	PVC	8	1.60	6.00	399.33	8.2	Clay	Street	1
752	69	VCP	8	0.35	9.95	399.17	5.7	Sand	Easement	3
753	58	VCP	10	0.28	11.87	399.00	5.5	Sand	Highway	3
754	58	VCP	8	0.52	5.29	398.99	5.4	Sand	Street	3
755	23	PVC	8	3.15	5.00	398.70	8.2	Clay	Street	1
756	70	RC	8	0.86	7.00	398.69	7.9	Clay	Street	1
757	37	PVC	8	0.50	10.00	398.67	7.9	Clay	Street	1
758	54	VCP	8	0.38	9.13	398.60	5.7	Sand	Alley	3
759	69	VCP	8	0.00	19.54	398.46	6.5	Sand	Easement	3
760	46	VCP	8	0.45	7.21	398.34	4.1	Sand	Street	3
761	67	RC	42	0.24	10.00	398.18	8.2	Loam	Street	5
762	8	PVC	8	0.40	15.00	398.04	7.9	Clay	Street	1
763	46	VCP	8	0.48	6.39	398.00	6.5	Sand	Street	3
764	57	VCP	8	0.48	6.39	397.99	6.5	Sand	Alley	3
765	19	PVC	8	2.61	5.00	397.68	8.2	Clay	Alley	4
766	25	PVC	8	2.05	8.00	397.59	8.2	Clay	Street	1
767	73	RC	18	0.01	5.00	397.53	8.2	Clay	Street	1
768	16	VCP	8	0.00	19.54	397.52	5.4	Sand	Highway	1
769	25	PVC	8	3.02	6.00	397.09	8.2	Clay	Easement	1
770	21	PVC	8	0.50	5.00	396.86	6.8	Sand	Street	1
771	65	RC	8	0.02	7.00	396.22	8.2	Clay	Alley	1
772	48	VCP	8	0.03	10.00	396.19	8.2	Loam	Street	1
773	56	VCP	8	0.39	8.85	396.09	5.7	Sand	Street	4
774	26	PVC	8	0.50	5.84	395.99	4.8	Sand	Easement	3
775	10	PVC	8	0.77	7.00	395.94	7.9	Clay	Alley	1

776	23	PVC	8	0.90	8.00	395.89	8.2	Clay	Alley	1
777	58	VCP	18	0.12	7.00	395.52	7.9	Clay	Street	1
778	45	VCP	8	0.28	11.87	394.99	5.4	Sand	Street	3
779	58	VCP	8	0.43	7.76	394.99	5.7	Sand	Easement	3
780	49	VCP	6	0.40	12.00	394.98	8.2	Loam	Street	2
781	22	PVC	8	3.36	7.00	394.51	8.2	Clay	Street	1
782	83	RC	6	2.80	7.00	394.50	8.2	Clay	Street	5
783	42	PVC	6	0.60	8.00	394.49	7.9	Clay	Alley	1
784	31	PVC	8	0.50	7.00	394.48	8.2	Clay	Easement	3
785	58	RC	6	1.80	8.00	394.43	8.2	Clay	Highway	3
786	58	VCP	8	0.42	8.03	394.31	5.5	Sand	Street	3
787	20	PVC	8	4.00	5.00	393.99	8.2	Loam	Street	1
788	10	PVC	8	5.10	5.00	393.82	8.2	Rock	Street	1
789	42	PVC	8	0.40	8.00	393.47	7.9	Clay	Street	1
790	19	PVC	10	1.18	5.00	393.24	7.9	Clay	Street	1
791	56	VCP	8	0.32	10.77	393.04	5.8	Sand	Easement	3
792	13	PVC	8	0.56	8.00	392.96	6.5	Sand	Street	1
793	46	VCP	8	0.69	0.64	392.84	5.7	Sand	Street	3
794	65	RC	30	0.38	7.00	392.67	8.2	Loam	Easement	3
795	24	VCP	12	0.00	19.54	392.53	4.8	Sand	Highway	3
796	58	VCP	8	0.39	8.85	392.49	5.7	Sand	Street	3
797	32	PVC	8	0.40	7.00	392.35	7.9	Clay	Street	1
798	24	PVC	8	1.01	7.00	392.09	8.2	Clay	Street	1
799	19	PVC	8	0.85	5.00	391.10	7.9	Clay	Street	1
800	59	VCP	30	0.02	18.99	390.80	5.7	Gravel	Street	1
801	75	VCP	8	0.40	10.00	390.68	8.2	Clay	Street	3
802	58	VCP	8	0.31	11.05	390.31	5.4	Sand	Alley	3
803	40	PVC	6	1.60	10.00	390.24	7.9	Clay	Street	1
804	63	VCP	8	0.30	7.00	390.22	7.5	Clay	Street	3
805	57	VCP	8	0.48	6.39	390.10	5.4	Sand	Street	3
806	14	PVC	8	0.30	7.00	390.09	7.9	Clay	Street	1
807	45	VCP	8	0.39	8.85	390.00	5.4	Sand	Easement	3
808	49	VCP	8	0.15	15.43	389.99	4.8	Sand	Street	1
809	14	PVC	24	0.14	7.00	389.92	8.2	Loam	Street	1
810	57	VCP	8	0.39	8.85	389.89	4.8	Sand	Street	3
811	22	PVC	10	0.28	11.00	389.81	7.9	Clay	Street	1
812	76	RC	8	1.28	5.00	389.64	8.2	Rock	Street	5
813	18	PVC	8	0.55	7.00	389.62	8.2	Clay	Street	1
814	14	PVC	8	0.40	7.00	389.07	7.5	Loam	Street	1
815	35	PVC	21	0.50	15.00	388.99	7.5	Clay	Street	1
816	46	VCP	8	0.31	11.05	388.29	5.5	Sand	Street	3

817	61	RC	12	0.72	10.00	388.05	7.9	Clay	Street	3
818	57	VCP	8	0.42	8.03	387.90	4.8	Sand	Alley	3
819	58	RC	12	1.00	5.00	387.75	6.8	Sand	Street	3
820	55	VCP	8	0.20	5.00	387.51	6.8	Sand	Street	1
821	28	PVC	10	0.30	6.00	387.00	7.9	Clay	Street	1
822	17	PVC	8	0.36	7.00	386.40	7.9	Clay	Street	1
823	46	VCP	8	0.31	11.05	386.19	5.4	Sand	Street	3
824	91	RC	33	0.72	8.00	385.89	8.2	Loam	Easement	3
825	101	VCP	8	0.30	8.00	385.54	8.2	Clay	Street	5
826	95	VCP	6	2.18	7.00	385.54	8.2	Clay	Easement	5
827	60	RC	18	0.50	7.00	385.26	8.2	Clay	Easement	3
828	28	PVC	8	0.50	5.00	385.17	8.2	Clay	Highway	1
829	32	PVC	8	2.46	8.00	384.67	8.2	Rock	Street	1
830	60	VCP	24	0.12	6.00	384.61	6.5	Sand	Street	3
831	22	PVC	8	0.33	10.00	384.40	6.7	Sand	Street	1
832	67	VCP	10	0.30	7.00	384.39	8.2	Clay	Street	1
833	57	VCP	8	0.42	8.03	383.80	4.1	Sand	Street	4
834	14	PVC	24	0.51	5.00	383.62	8.2	Loam	Easement	1
835	32	PVC	8	0.26	8.00	383.54	7.9	Clay	Street	1
836	22	PVC	8	0.70	7.00	383.42	6.8	Sand	Easement	1
837	65	RC	24	0.72	15.00	383.37	7.9	Clay	Street	5
838	57	VCP	8	0.41	8.31	383.09	5.7	Sand	Street	3
839	67	VCP	8	0.94	5.00	382.89	7.5	Clay	Street	5
840	61	RC	36	0.03	12.00	381.86	7.9	Clay	Street	1
841	60	RC	36	0.07	10.00	381.83	8.2	Loam	Street	1
842	64	VCP	8	0.40	7.00	381.42	7.9	Clay	Street	1
843	21	PVC	15	0.15	8.00	381.39	7.9	Clay	Street	1
844	55	VCP	8	0.40	8.58	381.35	4.8	Sand	Easement	3
845	60	RC	36	0.07	10.00	381.32	8.2	Loam	Alley	1
846	50	VCP	8	1.20	5.00	380.90	8.2	Clay	Highway	1
847	35	PVC	8	0.30	8.00	380.88	8.2	Clay	Alley	5
848	57	VCP	24	0.18	14.61	380.71	4.8	Gravel	Easement	4
849	66	VCP	8	0.24	12.96	380.27	4.8	Sand	Street	3
850	37	PVC	15	0.30	10.00	380.24	7.5	Loam	Highway	1
851	55	VCP	10	0.20	7.00	380.21	6.8	Sand	Highway	2
852	53	VCP	8	0.37	9.40	380.00	5.4	Sand	Street	3
853	69	RC	12	0.20	5.00	379.84	8.2	Clay	Street	1
854	62	VCP	8	0.35	6.00	379.67	7.5	Loam	Street	3
855	20	PVC	8	1.15	8.00	379.40	8.2	Clay	Easement	1
856	59	VCP	18	0.40	12.00	379.03	7.9	Clay	Street	1
857	11	PVC	8	2.40	7.00	379.03	7.9	Clay	Street	1

858	63	VCP	10	0.14	5.00	378.99	7.9	Clay	Street	2
859	45	PVC	8	0.30	5.00	378.95	8.2	Loam	Easement	2
860	65	RC	12	0.50	5.00	378.87	7.9	Clay	Highway	3
861	60	RC	6	0.30	12.00	378.79	7.9	Clay	Street	1
862	54	VCP	8	0.63	2.28	378.36	6.5	Sand	Easement	1
863	16	PVC	8	1.19	5.00	377.18	8.2	Clay	Street	1
864	17	PVC	10	0.47	8.00	376.89	7.9	Clay	Street	1
865	16	PVC	8	3.80	7.00	376.57	8.2	Rock	Street	1
866	51	RC	6	0.70	7.00	376.48	7.9	Clay	Street	1
867	60	RC	15	0.04	10.00	375.90	8.2	Rock	Easement	1
868	13	PVC	15	0.65	8.00	375.89	7.9	Clay	Street	1
869	72	VCP	8	0.59	3.38	375.70	4.8	Sand	Street	1
870	18	PVC	8	1.12	7.00	375.52	7.9	Clay	Street	1
871	18	RC	66	0.03	7.00	375.49	7.9	Clay	Easement	1
872	32	PVC	8	0.34	10.22	375.30	5.4	Sand	Street	3
873	60	VCP	8	0.42	8.03	375.07	4.8	Sand	Street	3
874	16	PVC	8	2.84	7.00	375.05	8.2	Clay	Street	1
875	50	VCP	8	0.38	9.13	374.99	6.5	Sand	Street	3
876	40	PVC	12	0.80	12.00	374.92	7.9	Clay	Street	1
877	18	PVC	8	0.80	7.00	374.71	8.2	Clay	Street	1
878	15	PVC	8	3.59	5.00	374.70	8.2	Loam	Alley	1
879	39	VCP	8	0.68	0.91	374.38	5.7	Sand	Street	1
880	13	PVC	16	0.26	5.00	374.21	8.2	Clay	Street	1
881	47	VCP	10	0.81	10.00	374.18	7.9	Clay	Alley	1
882	12	PVC	8	0.50	9.00	373.66	7.9	Clay	Easement	1
883	17	PVC	8	0.40	7.00	373.65	8.2	Clay	Highway	1
884	64	VCP	8	0.39	8.85	373.59	4.8	Sand	Easement	3
885	38	RC	27	0.20	10.00	373.55	8.2	Loam	Street	3
886	18	PVC	15	0.15	10.00	373.40	8.2	Loam	Street	1
887	54	VCP	15	0.30	10.00	372.95	7.9	Clay	Street	1
888	57	VCP	12	0.01	5.00	372.85	8.2	Clay	Street	1
889	46	VCP	8	0.38	9.13	372.85	6.5	Sand	Easement	3
890	22	PVC	8	3.10	8.00	372.08	8.2	Rock	Street	1
891	53	VCP	8	0.56	4.20	371.99	4.8	Sand	Street	3
892	10	PVC	8	8.00	6.00	371.97	8.2	Rock	Street	1
893	18	PVC	8	0.40	8.00	371.63	8.2	Clay	Street	1
894	41	PVC	8	0.40	6.00	371.54	5.8	Sand	Easement	1
895	31	PVC	18	0.32	8.00	371.48	7.9	Clay	Easement	1
896	22	PVC	12	0.50	7.00	371.13	7.9	Clay	Street	1
897	62	RC	27	1.00	5.00	371.06	8.2	Clay	Street	5
898	23	PVC	8	4.29	6.00	370.56	8.2	Loam	Street	1

899	28	PVC	10	0.30	10.00	370.48	7.9	Clay	Street	1
900	47	VCP	6	0.50	7.00	370.27	5.3	Sand	Street	1
901	63	RC	6	1.30	8.00	370.22	8.2	Loam	Easement	2
902	67	VCP	8	3.35	7.00	370.08	8.2	Clay	Street	1
903	58	VCP	8	0.37	9.40	370.00	4.8	Sand	Street	3
904	57	VCP	8	0.11	16.53	369.80	5.7	Sand	Street	3
905	59	RC	10	0.20	5.00	369.79	7.9	Clay	Street	3
906	79	VCP	6	0.60	7.00	369.53	8.2	Clay	Highway	3
907	58	VCP	8	0.42	8.03	368.99	4.8	Sand	Easement	3
908	13	PVC	8	0.70	7.00	368.70	7.9	Clay	Highway	1
909	63	RC	8	0.02	12.00	368.38	7.9	Clay	Street	3
910	37	PVC	10	0.80	15.00	368.19	7.9	Clay	Street	1
911	21	PVC	8	0.39	8.85	368.04	5.4	Sand	Street	3
912	29	PVC	10	0.80	16.00	367.12	7.9	Clay	Alley	1
913	18	PVC	8	1.30	10.00	367.07	8.2	Clay	Street	1
914	75	RC	8	0.50	8.00	366.87	8.2	Clay	Street	1
915	54	VCP	15	0.32	7.00	366.48	7.9	Clay	Street	1
916	22	PVC	8	0.38	7.00	366.19	7.9	Clay	Alley	1
917	46	VCP	8	0.32	10.77	365.99	6.9	Sand	Street	3
918	39	VCP	8	0.34	10.22	365.99	5.5	Sand	Street	3
919	65	VCP	12	0.50	5.00	365.97	7.9	Clay	Street	3
920	40	PVC	12	0.34	8.00	365.91	7.9	Clay	Street	1
921	25	PVC	8	0.24	10.00	365.86	7.9	Clay	Street	1
922	16	PVC	8	0.75	7.00	365.80	8.2	Rock	Easement	1
923	16	PVC	8	0.38	5.00	365.76	7.9	Clay	Street	1
924	38	RC	27	0.20	10.00	365.08	8.2	Loam	Street	3
925	59	VCP	18	0.40	12.00	365.07	7.9	Clay	Street	1
926	37	VCP	18	0.21	13.79	365.06	5.8	Gravel	Alley	3
927	18	PVC	8	0.33	7.00	364.86	7.9	Clay	Street	1
928	65	VCP	15	0.52	5.29	364.80	5.4	Sand	Street	3
929	58	VCP	6	0.70	5.00	364.08	8.2	Clay	Street	1
930	29	PVC	8	0.38	7.00	364.04	8.2	Clay	Easement	1
931	23	PVC	8	2.80	5.00	363.93	8.2	Clay	Alley	1
932	75	RC	15	0.15	7.00	363.76	7.5	Loam	Street	1
933	14	PVC	8	0.50	8.00	363.76	7.9	Clay	Street	1
934	58	VCP	8	0.40	8.58	363.75	5.7	Sand	Alley	3
935	15	PVC	8	0.78	7.00	363.68	7.9	Clay	Street	1
936	22	PVC	12	0.46	7.00	363.36	7.9	Clay	Street	1
937	24	PVC	8	1.20	10.00	363.18	8.2	Clay	Street	1
938	41	PVC	15	1.40	5.00	363.17	8.2	Loam	Street	1
939	43	RC	27	0.00	7.00	362.85	8.2	Clay	Street	1

940	19	PVC	8	0.35	7.00	362.53	8.2	Clay	Street	1
941	20	PVC	8	0.87	12.00	361.75	7.9	Clay	Street	1
942	30	PVC	8	0.30	11.32	361.70	5.4	Sand	Easement	3
943	57	VCP	8	0.42	8.03	361.60	5.7	Sand	Street	5
944	37	VCP	21	0.18	14.61	361.54	4.8	Gravel	Street	1
945	50	VCP	8	0.39	8.85	361.54	4.8	Sand	Street	3
946	41	VCP	8	0.31	11.05	361.22	4.8	Sand	Alley	1
947	69	RC	8	1.34	8.00	361.11	7.9	Clay	Street	1
948	43	PVC	6	1.00	10.00	360.87	8.2	Clay	Easement	1
949	58	VCP	8	0.01	10.00	360.83	8.2	Clay	Street	1
950	54	VCP	15	0.30	10.00	360.76	7.9	Clay	Street	1
951	65	VCP	8	1.20	13.00	360.59	7.9	Clay	Alley	3
952	49	VCP	8	1.28	6.00	360.37	6.7	Sand	Street	1
953	20	PVC	8	1.50	5.00	360.33	8.2	Clay	Alley	1
954	71	RC	8	0.33	7.00	360.32	7.9	Clay	Street	3
955	60	RC	10	0.01	7.00	360.28	8.2	Clay	Street	1
956	18	PVC	8	0.60	10.00	360.22	6.85	Loam	Street	1
957	13	PVC	15	0.20	7.00	360.22	7.9	Clay	Street	1
958	23	PVC	8	1.62	10.00	360.06	8.2	Clay	Street	1
959	45	VCP	8	0.39	8.85	360.00	5.4	Sand	Easement	3
960	45	VCP	8	0.39	8.85	360.00	4.8	Sand	Street	3
961	36	VCP	8	0.40	8.58	360.00	5.4	Sand	Street	3
962	46	VCP	8	0.50	5.84	359.99	5.4	Sand	Alley	3
963	83	RC	36	0.26	7.00	359.83	7.9	Clay	Alley	1
964	29	PVC	18	0.74	8.00	359.82	7.9	Clay	Street	1
965	59	VCP	10	0.50	7.00	359.80	8.2	Rock	Street	3
966	40	PVC	10	0.30	7.00	359.53	8.2	Loam	Street	1
967	6	PVC	8	0.14	15.70	359.45	5.4	Sand	Street	3
968	50	VCP	8	0.36	9.68	359.22	6.5	Sand	Easement	3
969	19	PVC	10	0.38	7.00	358.85	8.2	Clay	Street	1
970	71	RC	42	0.04	6.00	358.81	7.9	Clay	Easement	5
971	47	VCP	8	0.17	14.88	358.59	4.8	Sand	Highway	3
972	64	RC	10	2.20	10.00	358.51	8.2	Loam	Street	3
973	25	PVC	8	3.75	8.00	358.41	8.2	Loam	Street	1
974	46	VCP	8	0.31	11.05	358.31	4.8	Sand	Street	3
975	71	VCP	10	0.22	10.00	358.06	6.8	Sand	Street	5
976	49	PVC	8	0.48	6.39	358.06	6.5	Sand	Alley	3
977	21	PVC	8	0.29	11.59	357.98	5.8	Sand	Street	3
978	71	VCP	8	0.44	7.48	357.94	6.9	Sand	Street	3
979	73	RC	27	0.48	5.00	357.84	8.2	Clay	Easement	3
980	84	VCP	12	1.10	5.00	357.74	8.2	Clay	Street	3

981	6	PVC	24	0.30	5.00	357.72	8.2	Loam	Alley	1
982	49	VCP	6	1.00	7.00	357.61	8.2	Clay	Alley	2
983	64	VCP	8	0.55	4.47	357.49	4.8	Sand	Highway	3
984	35	PVC	8	1.00	7.00	357.41	8.2	Clay	Street	1
985	65	RC	8	0.03	7.00	357.26	8.2	Clay	Street	1
986	58	VCP	8	0.59	3.38	356.99	5.5	Sand	Street	3
987	64	VCP	10	0.24	10.00	356.96	8.2	Rock	Highway	1
988	38	VCP	8	0.38	9.13	356.95	5.4	Sand	Street	1
989	21	PVC	8	0.42	8.03	356.94	5.4	Sand	Street	3
990	43	VCP	8	0.42	8.03	356.60	5.4	Sand	Highway	3
991	41	VCP	8	0.32	10.77	356.57	5.4	Sand	Alley	3
992	72	VCP	15	0.14	15.70	356.20	7.4	Sand	Alley	1
993	45	VCP	8	0.39	8.85	355.99	5.4	Sand	Street	3
994	71	VCP	8	0.30	11.32	355.90	5.7	Sand	Street	4
995	59	RC	12	0.64	13.00	355.75	8.2	Rock	Street	1
996	68	RC	48	0.20	10.00	355.75	8.2	Loam	Street	3
997	23	PVC	8	0.02	5.00	355.67	8.2	Clay	Street	1
998	29	PVC	48	0.20	8.00	355.65	8.2	Loam	Street	1
999	22	PVC	12	0.60	7.00	355.63	7.9	Clay	Street	1
1000	42	PVC	10	2.50	10.00	89.96	8.2	Loam	Street	1

Biographical Information

Salar Shirkhanloo graduated with a Bachelor of Science in Civil Engineering from Islamic Azad University of Qazvin, Iran, in 2009. He then managed a building design office in Qazvin until 2011. He graduated with his Master of Engineering in Geotechnical Engineering at the Sharif University of Technology, Tehran, Iran, in 2013. After graduation, Salar worked as a construction cost estimator for Nik-Pey Construction Company in Qazvin, Iran, from 2013 to 2015. In 2015, he joined Tahkim Bana Building Design Office in Qazvin and worked as a professional structural engineer by 2018. Meanwhile, Salar taught various courses in the civil engineering area as an adjunct professor at Islamic Azad University of Qazvin, Iran, and Kar University of Qazvin, Iran, from 2013 to 2018. Salar's interest in civil engineering made him decide to continue his studies at the doctorate level. He joined the Center for Underground Infrastructure Research and Education (CUIRE) at the University of Texas at Arlington (UTA) to work on his Doctoral studies under the supervision of Dr. Mohammad Najafi in Fall 2019. During his time at UTA/CUIRE, Salar was a Teaching Assistant for various graduate courses such as Soils and Foundation in Construction Management, Geotechnical Aspects of Construction, Construction Contracts and Specifications, Construction Materials and Methods, Introduction to Mechanics for Construction, and Construction Design. Great enthusiasm for construction asset management and data science led Salar to complete his dissertation on the evaluation of decision-making prediction models for sewer pipes asset management in summer 2022. At the same time Salar completed his PhD, he joined the University of North Texas as a Clinical Assistant Professor.