COMPARISON OF CONDITION PREDICTION MODELS TO PRIORITIZE SEWER PIPE INSPECTIONS

By

MADHURI ARJUN

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 2023

DEDICATION

This dissertation is dedicated to my husband, Arjun Nanjundappa, who has supported and encouraged me through all the challenges of graduate school and life. I appreciate having you in my life. This work is also for my son, Agastya Arjun, who has always given me peace and love during my tough times and helped me to achieve this goal.

``

ACKNOWLEDGMENTS

I would like to express my most sincere gratitude and appreciation to my academic advisor Dr. Mohammad Najafi, P.E., F. ASCE, 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. I feel honored to have assisted him in his research projects and publications over the last three years at CUIRE. He was always there to help me when I was in need, and I am indebted to the opportunities he provided me. Dr. Najafi has always taken the financial burden off my shoulders for paying tuition throughout my Ph.D. studies by always securing me the graduate teaching assistantship. My dream of pursuing and completing Ph.D. would not be possible without my supervisor Dr. Najafi.

I express my heartfelt thanks to Dr. Melanie Sattler, Dr. Jay Michael Rosenberger, Dr. Vinayak Kaushal, and Dr. Syed Qasim 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 thank my friend and great Ph.D. colleague Shah Rahman, Practice leader at KCI Technologies for his help and support in sewer pipe inspection data collection from various city munipalities. Without his professional connections, I would have definitely not been able to collect data from eleven city munipalities.

Again, I would like to extend my special thanks to my husband, Arjun Nanjundappa for his keen interest in my research study and endless support throughout the model development phases. Without his training and help, this study would not have been accomplished by me. I would be ungrateful if I forgot to thank my parents, who always supported and guided me to take the right steps and action in all phases of my life.

ABSTRACT

COMPARISON OF CONDITION PREDICTION MODELS TO PRIORITIZE SEWER PIPE INSPECTIONS

Madhuri Arjun, Ph.D.

The University of Texas at Arlington, 2023

Supervising Professor: Dr. Mohammad Najafi

Over time, wastewater collection systems deteriorate, necessitating ongoing adjustments and the development of asset management frameworks by utility proprietors to maintain the performance of their assets. Any asset management framework should emphasize the importance of asset inspection and condition assessment for system-efficient operation and maintenance. In the United States, closed-circuit television (CCTV) is the most common method for inspecting the interior of sewer pipelines. This procedure is expensive and time-consuming due to a city's extensive inventory of pipes. Due to the immense quantity of these pipes, every municipality can only inspect some sections of sanitary sewer pipes promptly.

Therefore, the main objective of the research is to develop models capable of predicting the future condition of wastewater pipelines. The results of the models can be used to evaluate the need for inspection, rehabilitation, and replacement of sanitary sewer pipes. This dissertation utilized sewer pipe inspection data from eleven utilities from four US regions, namely Southeast, Southcentral, Northeast and Midwest. This data set contains six independent variables, which includes pipe age, diameter, length, material, soil native type, and slope, and one dependent variable, sewer pipe condition rating, based on PACP scores ranging from 1 to 5. This study evaluated the oversampling technique to address the imbalanced dataset employing the SMOTE method. Several machine learning algorithms created prediction models, including Logistic Regressions, Decision trees, Random Forests, AdaBoost, Gradient Boosting, and XGBoost with default parameter, AdaBoost, Gradient Boosting trees and XGBoost with oversampled hyperparameter Tuned. The other objective of this dissertation is a comprehensive investigation of the effectiveness of various machine learning methods using a resampled dataset.

Numerous evaluation metrics-Accuracy, F1-score and area under the curve (AUC), were calculated to compare the efficacy of developed models. The XGBoost with hyperparameter Tuned model had the best performance scores for all the dataset under different US regions, while the multinomial logistic regression decision tree model had the lowest performance scores accuracy. It was determined that tree-based models performed better than other models and that hyperparameter tuning was more effective in boosting trees.

Note-Please check Appendix A for a list of abbreviations.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION AND BACKGROUND

1.1 Introduction

The U.S. underground sewer systems are a significant part of municipal infrastructure, comprising thousands of miles of pipelines designed to carry and transport domestic sewage and stormwater runoff to the treatment plants (Najafi & Gokhale, 2022). Most sewer pipes in the U.S. operate using a gravity-based system. Gravity sewer systems rely on the land's natural slope to transport wastewater from higher to lower elevations, directing it to a treatment facility or a disposal point. The United States has over 800,000 miles of publicly owned sewer pipelines and over 500,000 miles of privately owned sewer laterals. 240 million Americans have access to 14,748 wastewater treatment facilities. It is anticipated that 56 million additional people will use concentrated treatment facilities by 2032. (Malek Mohammadi, 2019).

According to Environmental Protection Agency (EPA) research, up to half of the buried assets in investigated systems may be past the midpoint of their service lives (EPA, 2007). Most municipal sewers are a significant portion of the wastewater infrastructure in the United States is over a century old, aging, chemical, and environmental variables all result in at least 23,000 to 75,000 sanitary sewer overflows annually (EPA, 2015).

On the most recent infrastructure report card, published by American Society of Civil Engineering (ASCE) in 2021, the wastewater infrastructure received a "D plus" grade. According to ASCE, water and wastewater systems in the United States are readily aging, and an investment deficit of $150 billion must be addressed by 2025 to keep up with the needs (ASCE, 2021). Besides, the population of the U.S. is expanding and changing geographically. This demands investment in new infrastructure and maintenance of existing infrastructure in places with declining populations and confined budgets (EPA, 2009).

According to AWWA (2012), various municipalities and agencies prioritize sewage restoration rather than adding new sewer lines to meet growth or upgrading treatment plants. Inadequate maintenance and poor asset management methods raise the danger of inflow and infiltration, sanitary sewage overflows and sinkholes. Failure to manage clean sewer systems could harm human health while leading to expensive property damage and emergency repairs (Kumar et al., 2018).

In contrast to reliant maintenance procedures used by certain municipalities after pipe breakdown, preventive maintenance should involve inspection and maintenance activities before failure or irreversible deterioration (Fenner, 2000). Table 1-1 presents many factors that could lead to sanitary sewer pipe deterioration. Pipeline deterioration causes are typically categorized as one of the below that follow:

- Structural – cracks, fractures, breaks, and so on
- Hydrostatic – flooding, encrustation, and grease
- Corrosion – chemical and external corrosion
- Erosion
- Operational problems – roots, blockages, debris, and so on

Table 1-1 Factors Known to Influence Sanitary Sewer Deterioration
(Davies et al. 2001)

Construction factors	External Factors	Other Factors
Installation method	Surface use	Sewage characteristics
Standard of artistry	Surface loading (including construction traffic)	Use of appropriate maintenance
Sewer size	Surface type	Asset age
Sewer depth	Traffic characteristics	Sediment level
Sewer pipe material	Water primary bursts/leakage	Surcharge
Bedding material and type	Ground movement	
Joint type and material	Maintenance of other buried services	
Pipe section length	Groundwater level	
Connections	Infiltration/exfiltration	
	Soil/backfill type	

As previously stated, sewer pipes are a vital component of wastewater systems since they connect points of wastewater generation to treatment plants. There may be a decline in structural and operational efficiency as sewer systems age. Old or deteriorated pipes can cause problems for people's health, the environment, and the economy (Opila, 2011).

Maintenance and rehabilitation methods are essential in sustaining the pipeline’s operation at an acceptable level of service and offering cost-effective ways to avoid unplanned failures. Previously, sewer pipe repair or rehabilitation was only done when a pipe collapsed or failed. The current trend, however, is to repair and manage pipe systems before they collapse. Municipalities and utilities have begun to adopt asset management systems to attain this goal. Infrastructure asset management is a thorough and efficient process. Method for keeping pipeline systems in good working order. An efficient asset management strategy can include several techniques to assist utility owners and municipalities understand the time and related costs of deteriorating pipe repair, rehabilitation, or replacement (Loganathan, 2021).

Sewer pipeline deterioration depends on many factors and steps, making it harder for municipalities to locate collapse-prone pipes. In recent years, sewage pipeline inspection and monitoring have intensified to prevent further collapse and failure. Hence, pipe deterioration models that predict sewer pipeline conditions must be developed. This dissertation examines statistical and A.I. algorithms for sanitary sewage pipe condition prediction. Clean sewer pipe effect aspects will also be discussed.

1.2 Need Statement

Several eminent researchers in the United States have developed condition prediction models to identify the critical factors that influence the deterioration of sanitary sewer pipelines. The developed deterioration models utilize statistical methods and AI-based algorithms. However, a single standard model has yet to be created by collecting data from different geographical areas in the United States. This proves that city municipalities cannot employ the prediction models developed by past researchers to prioritize inspection operations on sewer pipes (Shirkhanloo, 2022). One of the most critical limitations of current sewer prediction models has been the need for data from different geographical locations to train and validate reliable models. Several contributors suggested improving sewer pipe condition prediction models from various perspectives, which includes below factors:

- Najafi and Kulandaivel (2005) said that the neural network model for sewer pipe deterioration could be improved by adding more historical input variables, such as surface load, groundwater, bedding conditions, soil corrosion, stability, and sewer location.
- Chughtai (2008) recommended incorporating more variables like Soil type and its conditions. predicting the sewer pipe deterioration models. Future research should investigate the application of further prediction models.
- Mashford et al. (2011) suggested that pipe length and depth data must be incorporated as independent variables in developing prediction models.
- Sousa et al. (2014) proposed using higher-level deterioration models and comparing the results to machine learning and neural network models.
- Kabir et al. (2018) presented that the developed sewer structural condition prediction models may be improved by assessing the effects of additional independent variables, such as sewer function, groundwater level, soil type, road class, and initial quality of construction.
- Mohammadi (2019) stated that a prediction model must be able to predict each of the five condition levels independently instead of transforming them into binary classes.
- Loganathan (2021) developed the prediction models by adopting advanced machine learning algorithms using data collected from one city municipality. The research suggested that models must be validated with the inspection data collected from different cities.
- Shirkhanloo (2022) used supervised learning algorithms to develop prediction models by gathering data from a single municipality. The researcher pointed out that the inspection data from more cities with variables like pipe diameter, material, length, depth, soil, and PACP ratings are required to develop the prediction models that the city municipalities can employ to prioritize the sewer pipe inspection.

From the studies mentioned earlier and the developed models, it is shown that there is a significant knowledge gap in identifying the vital variables affecting the deterioration of sewer lines. It was also presented that most studies needed more variations in the data collected from city municipalities and were found to be restricted. This is a significant limitation of the prediction models developed above as they are based on the single city analysis.

1.3 Objectives

The primary objectives of this research work include:

- To identify critical variables affecting sanitary sewer pipeline conditions.
- To develop an A.I.–based prediction model capable of forecasting the deterioration of sewer pipes.
- To compare major condition prediction models.
- To compare critical variables.

1.4 Scope of Research

The scope of this research is restricted to the study of sanitary sewers with gravity flow using the PACP scores that the operators' record by carrying out the Closed-circuit television (CCTV) inspection for modeling the deterioration of pipe systems. No pipe rehabilitation methods details have been considered to maintain the consistency in the data collected. The condition of sewer pipes is categorized based on the Pipeline Assessment and Certification Program (PACP) developed by the National Association of Sewer Service Companies (NASSCO). Table 1-2 presents the scope of this dissertation.

Table 1-2 Scope of Research

Included	**Not Included**
Sanitary sewer pipes	Stormwater pipes
Gravity sewer pipes	Force main sewer pipes.
Inspected pipes based on PACP guidelines	Inspected pipes based on other guidelines
VCP, PVC, RCP, UnReinCONC, RPM, DI, CI, AC, HDPE, PCCP, FRP, CLC, CMP.	Other not included Sewer pipes
Sanitary sewer pipes without any repair or rehabilitation history	Pipe segments that have a history of pipe maintenance

1.5 Research Methodology

The artificial intelligence (A.I.) models developed in this dissertation are used to predict the condition rating of individual sewer lines by considering the physical features of the pipelines and various environmental factors. These significant variables may lead to the eventual deterioration of sewer pipelines. As illustrated in Figure 1-1, the following steps are followed in this methodology to achieve the intended result of the research.

Step 1: Problem Statement

Step 2: Define Objectives and Scope of Work

Step 3: Comprehensive Literature Review

Step 4: Data Collection and Data Preparation

Step 5: Development of Deteriorating Models.

Step 6: Model validation.

Step 7: Comparing A.I. Models Performances

Step 8: Best Model Selection

Step 9: Development of Feature Importance (Independent Critical Variables Coefficient Determination)

Step 10: Model Performance Validation (with Identified Critical Variables only)

Step 11: Identify Critical Variables

Step 12: Conclusions

Step 13: Recommendations for Future Research

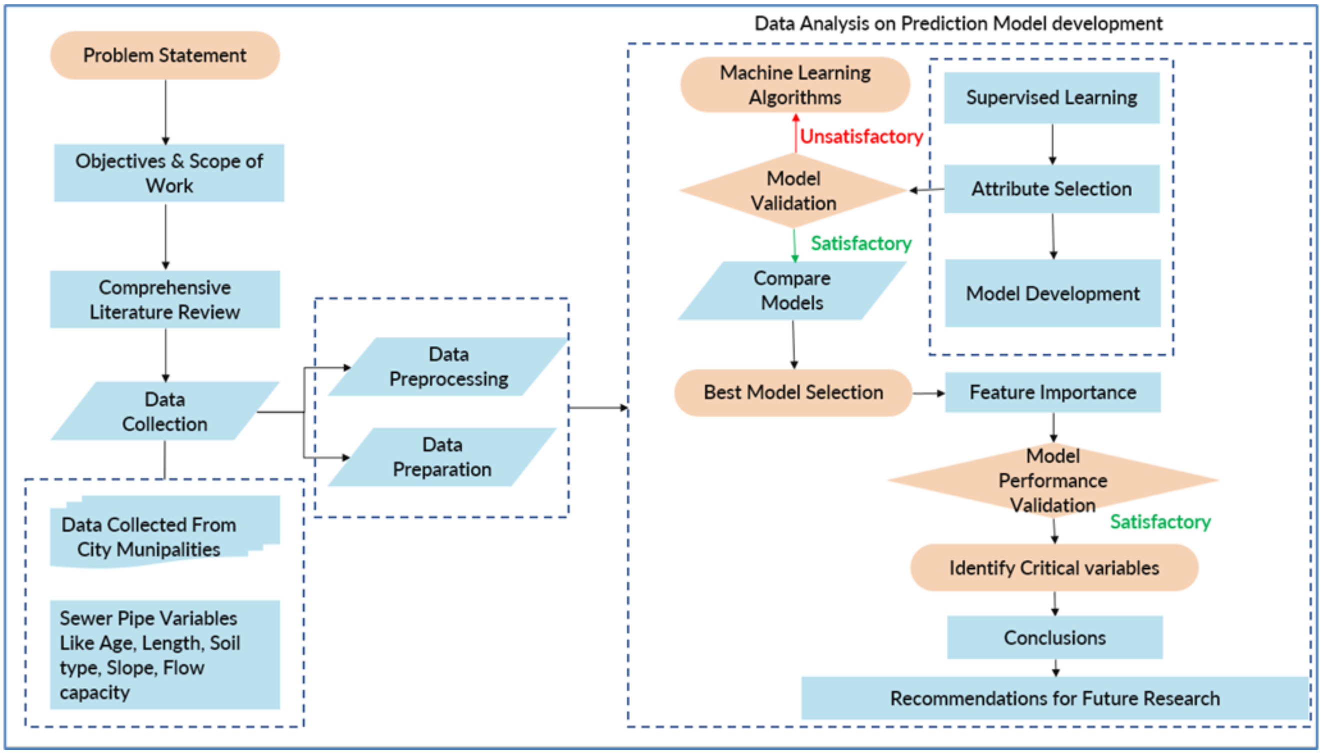

Problem Statement
Objectives & Scope of Work
Comprehensive Literature Review
Data Collection
Data Collected From City Munipalities
Sewer Pipe Variables Like Age, Length, Soil type, Slope, Flow capacity
Data Preprocessing
Data Preparation
Data Analysis on Prediction Model development
Machine Learning Algorithms
Unsatisfactory
Model Validation
Satisfactory
Compare Models
Best Model Selection
Supervised Learning
Attribute Selection
Model Development
Feature Importance
Model Performance Validation
Satisfactory
Identify Critical variables
Conclusions
Recommendations for Future Research

Figure 1-1 Research Methodology

1.6 Expected Results

AI-based logistic regression decision tree models and supervised learning algorithms, such as bagging (Random Forest) and boosting (AdaBoost, Gradient Boost and XGBoost) methods are developed in this research to evaluate the deterioration of sewer pipes. The expected results of this dissertation are discussed below:

- A tool for sewer asset managers to make strategic decisions.
- A comprehensive comparison of the various methods that developers can use to select the optimal forecasting model.
- A condition prediction model was used for each case study to classify the sewer pipes into multiple classes.

1.7. Hypotheses

1.7.1 Hypotheses 1

Null hypotheses (H_0): Pipe material, diameter, age, slope, and depth variables do not influence sewer pipes' PACP score.

Alternative hypotheses (H_A): Pipe material, diameter, age, slope, and depth variables influence sewer pipes' PACP score.

1.7.2 Hypotheses 2

Null hypothesis (H_0): The tree-based models are expected to perform better than the other AI-based algorithms.

Alternative hypothesis (H_A): The tree-based models are not expected to perform better than the other AI-based algorithms.

1.8 Chapter Summary

Chapter 1 provided background information on the status of sanitary sewage pipes, significance of sewer inspection and maintenance procedures. This chapter also discussed the research needs, statement, objectives, scope of research, methodology, expected results, and hypotheses.

CHAPTER 2

LITERATURE REVIEW

2.1 History and Overview

According to the ASCE report card 2021, the public sanitary sewage pipelines span over 800,000 miles, and lateral sewers run around 500,000 miles, contributing to the important portion of the underground utilities and infrastructure (ASCE 2021). The ASCE report stated that the combined capital requirements for water and wastewater systems are forecasted to be $150 billion from 2016 to 2025, with a $105 billion investment gap between estimated and required funding (ASCE 2021). This identified investment gap highlights the importance of using the available budget in the most effective and efficient manner by choosing the proper asset management.

For sanitary sewer systems, asset management was first applied in the early 2000s, and the Environmental Protection Agency (EPA) acted significantly in supporting and formulating the principles on it (Syachrani, 2010). "Asset management can be described as handling infrastructure assets in order to minimize the overall cost of ownership and operation while offering the service levels customers desire" (EPA 2002).

The fundamental components of an asset management system include the identification, location, and condition of assets. Pipeline condition evaluation gives vital information about the physical and operational state of pipes, allowing for the estimation of remaining service life and long-term performance of infrastructure pipe systems. Pipe condition assessment can be calculated using standard coding systems and information gathered during the inspection procedure (EPA, 2009). Pipeline condition assessment provides essential details about the physical and functional state of pipes, which facilitates estimation of remaining service life and long-term performance of infrastructure pipe systems. Pipe condition assessment can be calculated using standard coding procedures, and data gathered during the inspection procedure can be utilized to develop prediction models (EPA, 2009). The condition assessment and prediction model results help municipalities and utility agencies to formulate a decision-making strategy for the asset's current and future state. This, in turn, guides the government to prioritize the assets which may need future investment.

2.2 The United States Sanitary Sewer System

In the mid-seventeenth century, there was no proper carrier system in place to collect raw sewage. The lack of sewage infrastructure did not provide a sanitary challenge at the time due to the low population density. But as the United States population began to increase in the early 1800s, managing sewage disposal became difficult, and hence arose the demand for developing sanitary sewer systems (Burian et al., 2000). The communities started developing sewer systems by adopting the expertise and methods which were popular in Europe and Asia to protect, promote public health and safety from sewage flooding.

The sanitary sewer system built initially was non-efficient in handling the drainage flow, which led to polluting the soil and groundwater. Occasionally contaminated drinking water results in disease epidemics. Therefore, the US municipalities review the situation and aim to find an alternative solution to design a comprehensive sanitary sewer system to solve sewage flooding and pollution by employing skilled engineers.

The comprehensive sanitary sewer system developed by the city municipalities comprises Combined Sewer Systems (CSS), Separate Sanitary Sewer and Storm Sewer Systems (SSS) (EPA, 2004).

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 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) (Malek Mohammadi 2019).

2.3 Asset management

"Asset management can be described as the management of infrastructural assets in order to minimize the overall costs for ownership and maintenance while providing the service levels that consumers desire. (EPA, 2002)."

New York's Department of Environmental Conservation (DEC) created a Municipal Sewage System Asset Management (MSSAM) to manage the sewer pipelines, and it defines Asset Management as “a system to achieve sewer pipelines optimal performance and longevity by performing the regular maintenance, upgrade to minimize disruptions, limit environmental impacts, and maximize sewer system management cost-effectiveness” (MSSAM guide 2015). In the United States, researchers and governments define asset management differently. Most asset management strategies incorporate inventory, essential asset prioritization, and financial planning to maintain performance. Aging sewer systems make asset management more critical. Urban drainage system data is undervalued compared to data-intensive fields such as bioinformatics and medical sciences (Tscheikner-Gratl et al., 2020). An agency's sewage asset data must be saved, processed so operators and decision-makers may use the asset data.

2.4 Sewer Pipes Deterioration Mechanisms and their Affecting Factors

Sanitary sewer pipeline systems are among the most capital-intensive infrastructure systems because of their direct and indirect effects on their environmental and financial surroundings (Najafi and Gokhale, 2005). The research by Davies et al. (2001) presented that the fundamental performance requirements for sewer operation are as follows:

1. Pipeline networks cannot have obstructions;
2. Sewage treatment plants must have adequate capacity; and
3. Sewage treatment plants.

2.4.1 Pipe Age

Pipe age is commonly referred to as the difference between the year of installation and the date of inspection. Pipe age commences upon installation (Kulandaivel, 2004). Numerous studies have demonstrated that the age of sewer pipelines significantly affects their condition (Ariaratnam et al. 2001, Chughtai 2008, Ana et al. 2009, Salman and Salem 2012, Laakso et al. 2018). As depicted in Figure 2-1, the serviceability of pipelines declines over time and is divided into five distinct stages (Misiunas, 2005). According to Singh and Adachi (2013), pipe failure is depicted by a bathtub curve, which is created by plotting the pipe failure rate versus time. As shown in Figure 2-3, the bathtub curve includes three distinct phases. The first is the early life stage, which has a high failure rate and exhibits problems shortly after installation. Human factors, pipe damage during construction, installation, and inappropriate pipe material can all contribute to failures during this time limit.

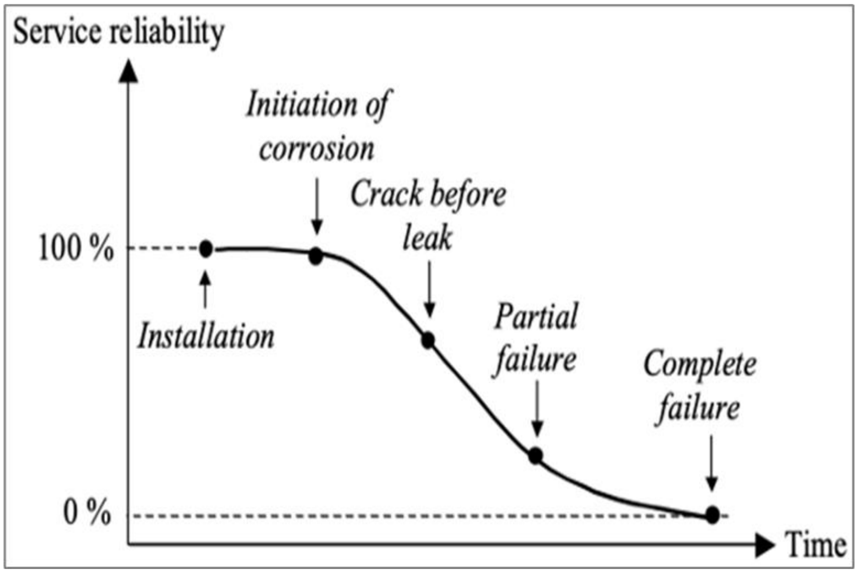


Figure 2-1 Serviceability of a Pipe
(Misiunas 2005)

The second phase represents the useful life of the conduit, with a failure rate that is extremely low and constant. Failures in the second phase could result from a number of unforeseen occurrences, such as exceptionally heavy cargo, earth movement, settlement, or third-party interference. Due to pipe deterioration over time due to aging, the third phase (wear-out life) has a high failure rate (Singh and Adachi, 2013). Contrarily, a small number of studies (Tafuri and Dzuray, 2000; Davies et al., 2001) concluded that age is not a significant factor in pipe deterioration.

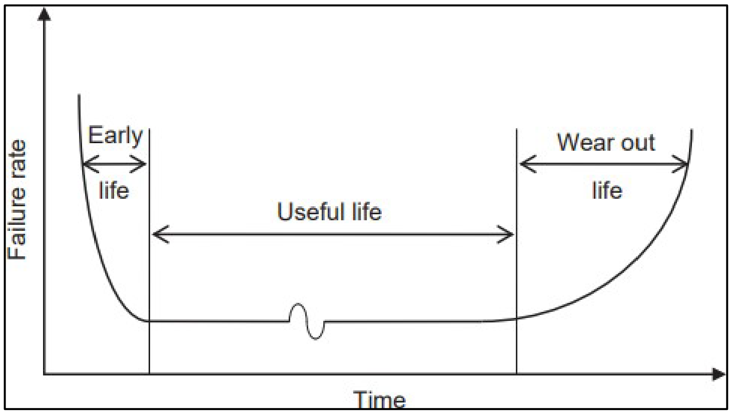


Figure 2-2 The Theoretical Bathtub Curve of Buried Pipe
(Singh and Adachi, 2013)

2.4.2 Pipe Material

Sewer pipes made of various materials respond differently to environmental factors such as soil type, water table, etc. (Salman, 2010). For instance, concrete pipelines are highly resistant to abrasion, clay pipes are highly acid-resistant. Plastic pipelines, like PVC or HDPE, are resistant to acidic and alkaline wastes, but they are susceptible to excessive deformations under load (Singh and Adachi, 2013). The material of the pipe can be used as an independent variable during the development of condition prediction models, and the results of the model can indicate whether this variable is significant or not. Davies et al. (2001) determined that pipe material is a significant variable and that there is a direct correlation between pipe material and the deterioration of sewer pipelines. Micevski et al. 2002 selected, using their Markov model, that concrete pipelines are more robust and long-lasting than clay pipes. Ana et al. 2009 indicated that concrete pipes performed better than masonry and clay pipes in the model. Pipes' manufacturing process is a contributor to their disparate aging characteristics. Typically, concrete pipelines are constructed in a controlled environment, resulting in high quality and durability. In contrast, masonry pipes are typically constructed on-site, and the quality of the pipes is affected by varying environmental conditions and shoddy craftsmanship.

In the model devised by Lubini and Fuamba (2011), pipe material was also significant. They discovered that reinforced concrete pipes are more resistant to deterioration than other pipes because reinforcing steel makes the conduit powerful enough to prevent structural damage. Bakry et al. (2016) demonstrated in their model that vitrified clay pipes performed better than asbestos cement and reinforced concrete pipes. Significant concert and polyethylene high-density pipelines were identified in the prediction model developed by Laakso et al. (2018). Deficiencies in the quality of certain samples of polyethylene high-density pipes were an explanation for the disparate behavior of pipe materials in their study. In contrast, Jeong et al. (2005) stated that the material of the conduit was not a significant variable in their study. According to their report, the class imbalance and limited number of data used to develop the prediction model could be a plausible explanation. In general, it would be easier to anticipate the deterioration behavior of pipes if distinct models were developed for each pipe material.

2.4.3 Pipe Diameter

Numerous studies have demonstrated that pipe size or pipe diameter is a significant factor in the deterioration process. When the diameter of a sewer pipe falls between 6 and 8 inches, it is classified as a smaller sewer pipe, and when it exceeds ten inches, it is classified as a larger sewer pipe. Based on condition prediction models developed in a few studies, it was determined that the rate of sewer pipe condition deterioration decreases as pipe diameter increases, whereas a few other studies indicate that larger-diameter pipes fail more frequently. Lubini and Fuamba (2011), Salman and Salem (2012), and Bakry et al. (2016) insisted that pipelines with a larger diameter perform better than those with a smaller diameter. Because larger-diameter pipelines can continue to function, albeit not necessarily at full capacity, when obstructions occur, whereas smaller-diameter pipes lose hydraulic flow. According to the study, larger-diameter pipelines are buried deeply, which may account for their superior structural condition. Therefore, larger-diameter pipelines have lower deterioration rates than smaller-diameter pipes (Malek Mohammadi et al. 2019, Micevski et al. 2002, Wirahadikusumah et al. 2001, and Najafi and Gokhale 2005).

In contrast, Tran et al. (2007) found that the conduit's size was insignificant. In addition, Jeong et al. (2005) found that larger pipes are more susceptible to deterioration because they have a greater surface area exposed to sewage and the adjacent soil.

2.4.4 Pipe Length

The length of a sewage conduit is measured between the entrance and exit manholes. According to Najafi and Gokhale (2022), shorter pipes deteriorate at a quicker rate than longer pipes due to the sharper bends along the length of longer pipes, which could result in less debris or obstructions. On the other hand, Malek Mohammadi (2019) indicated that longer sewer pipelines may have a higher rate of deterioration due to a higher flaw probability.

In addition, a few studies reveal a dual behavior in the condition of pipes in relation to variations in pipe length. According to Laakso et al. (2018), sewer pipes longer than 131 feet deteriorate more rapidly than other pipes in the network, while pipes shorter than 131 feet have almost no influence on the condition of the pipe. This consequence can be explained by the fact that longer pipes carry a greater danger of defects and bending stress. Moreover, lateral connections can result in structural damage, and lengthier pipelines contain more of them.

2.4.5 Pipe Slope

The slope of the sewer pipe is a significant factor in sanitary sewer pipe deterioration (Baur and Herz, 2002). The slope or gradient of a pipe can be estimated by dividing the difference between the elevations from the mean sea level (MSL) of the pipe at the beginning and the end to the inspection length, as illustrated in Equation 2-1.

$$\text{Slope} = \frac{Elevation\ at\ origin - Elevation\ at\ end}{Inspected\ Length} * 100 \quad \text{................................ Equation 2-1}$$

Evidence shows that flat sewer pipelines deteriorate more slowly than pipes with a steeper gradient. When the slope is steep, the flow rate will also be steep, making erosion easier (Najafi and Gokhale, 2005). It is claimed that pipelines with an extremely low gradient could facilitate sediment deposition, leading to clogging and obstructions. (Jeong et al., 2005) Sewer pipes with flat slopes tend to have lower velocities, causing wastewater to remain within the pipe for an extended period and resulting in the production of hydrogen sulfide naturally.

2.4.6 Pipe Depth

In contrast, pipe depth played no role in the prediction model devised by Davies et al. (2001). This is not to say that sewer depth has no effect on the deterioration of pipes when considered independently, but in data analysis based on the characteristics of pipe datasets, there may not be a direct relationship between pipe depth and sewer pipe condition level. According to Tran et al. (2007) and Ana et al. (2009), pipe depth was insignificant in their prediction models. Due to surface load, illegal connections, and tree root intrusion, generally speaking, shallowly buried pipelines would be subject to more defects and a higher rate of deterioration. In addition, increased cover depth above the pipelines reduces the impact of surface factors such as road traffic, road maintenance, and construction activities. Salman and Salem (2012) discovered the same outcome, and among the eight independent variables used in their model, pipe depth was the only insignificant variable. In their study, Laakso et al. (2018) found a correlation between installation depths between 6 and 10 feet and poor conditions, and they suggested a minimum installation depth of 5 feet due to winter cold.

To determine the appropriate depth of sewer pipelines, several factors, including soil type, water table, pipe material, pipe diameter, and regulations, must be considered. Diverse prediction models produce contradictory findings regarding the effect of depth on the deterioration of sewage pipelines. Khan et al. (2010) indicated that pipe depth is a significant variable in their prediction model and that any increase in depth has a negative effect on wastewater pipe condition levels. The rationale for this behavior may be the increased dead burden on the pipes and the increased likelihood of groundwater table.

2.4.7 Pipe Location / Surface Type

Obviously, the surface pressures above any underground utility structure will have an effect on it. The quantity of surface loading carried to the sewer pipe depends on the land use and the nature of traffic above the pipe. There is a correlation between the surface loading type and the sewer pipe (Kley and Caradot 2013, Najafi and Gokhale 2005), despite the fact that the frequency of surface loads makes it difficult to estimate their influence on deterioration. According to Bakry et al. (2016), sewage pipelines in proximity to industrial areas deteriorate more rapidly. Few studies (Tran et al. 2007, Micevski et al. 2002) concluded that the location of pipes has no significant impact on their structural integrity.

2.4.8 Pipe Soil Native Type

Different soil types react differently to pipe material, groundwater, and other pipe characteristics and environmental factors (Kaushal and Guleria, 2015). According to Wirahadikusumah et al. (2001), the underlying soil has a significant effect on sewage conduit deterioration. Comparing the condition of pipelines installed in stable versus unstable soil, Tafuri and Dzuray (2000) found that pipes installed in unstable soil experienced greater condition fluctuations. In addition, the type of soil surrounding a sewage pipe is one of the most important factors that can influence frost heave, soil-pipe interaction strength, and external corrosion, all of which can contribute to failure mechanisms (Najafi and Gokhale, 2022). When there is insufficient soil support around a sewage pipe, it may shift, leading to the formation of cavities that make the pipe more susceptible to deformation (Loganathan, 2021). In contrast, soil type was not a significant factor in the prediction model created by Laakso et al. (2018).

2.4.9 Corrosion

According to Shirkhanloo (2022), soil corrosivity is a soil property that increases the likelihood of external corrosion on pipe surfaces. Typically, corrosion in steel pipelines is caused by an electrochemical reaction between the pipe's exposed outer surface and the surrounding soil environment. There are varying degrees of corrosion resistance among conduit materials. Numerous variables, including soil acidity, resistivity, pH content, oxidation-reduction, sulfide, moisture, aeration, etc., have been observed to influence the corrosion rate (Loganathan, 2021). According to Najafi (2016), longitudinal failure can result from conduit wall deterioration caused by corrosion. Only a few studies have examined the effect of soil corrosivity on the deterioration of sewage pipelines; it should be noted.

2.4.10 Soil pH

Almost all studies in the field of subterranean corrosion (Wasim et al., 2018) indicate that the pH of the soil impacts the corrosion rate of buried pipelines. According to Najafi and Gokhale (2022), the pH of the soil is a useful indicator of external corrosion because different pH ranges result in various corrosion processes. There are three distinct pH ranges: alkaline (pH>7), neutral (pH=7), and acidic (pH7).

Hou et al. (2016) investigated the effect of soil pH on pipelines made from different materials. Cast iron pipelines are more likely to corrode than steel pipes under the same corrosive conditions, according to the findings.

2.4.11 Groundwater

Groundwater is the subterranean water found in soil, sand, and rock fissures and crevices. The presence of groundwater near or above sewer pipelines may cause water to flow through the conduit, thereby increasing structural defects, void formation, and support loss. In cohesive soil, an increase in the groundwater level may reduce the soil's cohesive strength and enlarge the void surrounding the conduit. Therefore, supporting soil can be readily washed away (loosened), and the pipe is more likely to collapse under these conditions. Typically, sewers located in areas subjected to elevated groundwater are at a greater risk of failure than sewers located in areas where the groundwater level is below the sewer level. According to Davies et al. (2001), the availability of groundwater around the conduit causes the loss of soil support and infiltration defect. In addition, the formation of voids and the absence of adequate support around the conduit contribute to sewer structural issues. Periodic water table in a cohesive soil may cause a decrease in soil strength and the potential for soil to be flushed into the sewer. Malek Mohammadi et al. (2019) determined that groundwater level is a significant variable based on a prediction model developed for the City of Tampa. They concluded that groundwater increases pipe burden and the risk of soil movement and infiltration. Typically, the groundwater level is not accounted for in pipeline inventories, and it has only been utilized as a variable in a handful of prediction models. The impact of groundwater level on the condition of sanitary sewer pipelines requires further investigation.

2.5 Condition Assessment for Sanitary Sewers

2.5.1 Introduction

Condition assessment is an essential component in infrastructure asset management, and it can be defined as the analysis of the data collected during the field inspection of sewers to evaluate their performance structurally and operationally (Loganathan, 2021). The asset's physical state can be assessed in the condition assessment procedure. Also, the deterioration pattern can be detected to predict their failure time. McDonald and Zhao 2001 presented an algorithm to carry out the condition assessment procedure.

of the existing sewers to calculate the numerical grade of the sewer asset, determining its structural and operation state. Figure 2-3 depicts the condition assessment algorithm.

2.5.2 Condition Scoring Methods of Sanitary Sewers.

The sewer condition assessment basic idea is to provide a comparison between the existing asset's structural and operational ability with that of a new asset (Shirkhanloo, 2022). There are various methodologies to develop the generic coding system on the sewer pipe state, and the most prominent methods in condition rating methods of sewer pipes are Water Research Centre (WRc) and Pipeline Assessment and Certification Program (PACP).

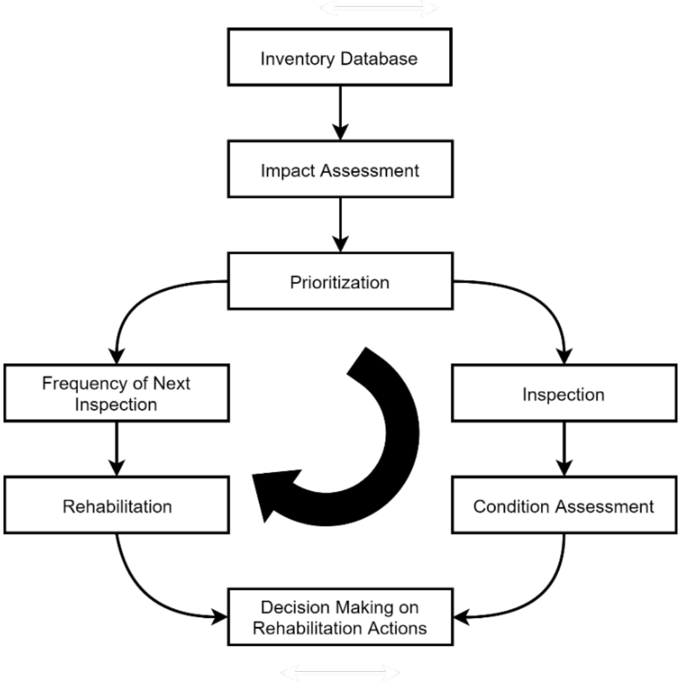


Figure 2-3 Condition Assessment Algorithm
(Adapted from McDonald and Zhao, 2001)

2.5.3 WRc Condition Scoring Method

A water research center is an institution in the United Kingdom devoted to studying diverse facets of water, such as its quality, availability, management, and environmental impact. Frequently, these institutes conduct research, provide education and training, and provide technical assistance to address water-related issues and develop sustainable solutions. In 1977, WRc developed a research project to design a generic coding system to assess sewer pipe conditions. During this research, in 1980, WRc published the world's first rehabilitation manual for sewers, which later in time became the standard for developing the protocols for the sewers (Chughtai and Zayed, 2001).

Individual scores are assigned based on evaluating these several factors, and an aggregate condition rating is determined for the sewer system. The condition rating can be expressed using a numerical scale or descriptive categories (such as excellent, good, average, and poor) (Thornhill and Wildbore, 2005).

The WRc condition scoring system may be modified or adapted by specific organizations or municipalities. Therefore, the precise details and scoring criteria may vary marginally depending on the context-specific guidelines followed (Opila, 2011).

2.5.4 PACP Condition Scoring Method

Pipeline Assessment and Certification Program (PACP) was developed in 2001 by the National Association of Sewer Service Companies (NASSCO) in association with WRc to design a standard for sewer condition assessment. PACP aims to construct a database to accurately identify, plan, prioritize, manage, and renovate sewer pipe assets based on condition assessment.

According to the NASSCO coding system, pipe defects and features can be classified into five categories. The defect classification includes classes for (1) continuous defects, (2) structural defects, (3) operational and maintenance, (4) construction features, and (5) other features (NASSCO, 2015).

Several factors, such as the significance of the defect, the extent of the damage, and the percentage of restriction to flow capacity or wall loss due to deterioration, are used to assign grades. The final condition rating is derived from the categories of structural, operation, and maintenance (O&M). Table 2-1 presents the steps and their respective definitions of the PACP condition rating representing the NASSCO 2015 manual. On a scale from 1 to 5, PACP ranks the condition of pipelines based on the results of CCTV inspections and operator evaluations. Condition 1 denotes that the pipe is in exceptional condition, whereas Condition 5 denotes that the pipe has failed or will soon fail. Piping with a condition rating of 5 must undergo immediate rehabilitation or replacement.

Table 2-1 PACP Condition Rating
(NASSCO, 2015)

Condition Grade	Description	Time to Failure
5-Immediate Attention	Defects requiring immediate attention	The 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 near future	The pipe will 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 soon

2.6 Prediction Models for Sanitary Sewer System

2.6.1 Significance of Condition Prediction for Sewers

Obviously, not all sewer pipelines in a collection would be in poor structural condition or be an imminent failure. Moreover, inspecting every sewer conduit in a system would be costly and time-consuming. As discussed in the previous section, the financial requirements for each inspection operation could be calculated based on the operation space and test setup complexity.

Consequently, it is necessary to identify the most crucial sewer pipes for inspection in the complete inventory. By predicting pipelines in poor condition in advance, reducing the frequency of sewer pipe inspections is possible. This pipe inspection prioritization would save any municipality thousands of dollars (Chae and Abraham 2001).

Predicting sanitary sewer pipe status is not new. Researchers have conducted many sewer pipe condition prediction studies using computer technology, machine learning algorithms, or artificial intelligence. Because municipalities store different data in their database inventory, there is no standard model. Thus, many towns need an asset management plan and inspection prioritizing (Loganathan, 2021).

2.6.2 Statistical Prediction Models

Statistical models use probabilistic historical data to characterize model output as a random variable. Based on historical data, statistical analyses employ "ideally suited approximations" (Wright et al. 2006). The random variables X that represents unknown quantities are statistical models. The parametric density function is used in statistical models, as stated by Dasu and Johnson (2003), to analyze mistakes and find probabilistic correlations between dependent and independent variables. Statistical models can more accurately predict the condition of sewage pipes than deterministic models, which produce quantitative results (Coles, 2011). Previous research utilized logistic regression, Markov chain, ordinal regression, and the cohort survival model to predict the condition of wastewater pipelines.

Ariaratnam et al. (2001) employed logistic regression to forecast sewer pipe condition states by considering pipe age, depth, material, diameter, and service kinds into account as independent variables. To determine the appropriate independent variables in the model, a linear regression variable selection method was applied. The Wald Test and likelihood-ratio test were used to determine the significance of the variables in this investigation. The likelihood-ratio test found that the model's important variables are pipe age, diameter, and sewer type. To validate the logistic regression model, a sensitivity analysis was undertaken. However, sensitivity analysis is insufficient to assess the efficiency of the logistic regression model.

Hahn et al. (2002) created an expert based on an expertise support system to prioritize sewer pipeline inspection. Interviews and case studies were used to build the Bayesian belief network model. Based on failure chances and repercussions, Sewer Cataloging, Retrieval and Prioritization System (SCRAPS) was created as a decision assistance tool. WRc's 1986 pipe assessment paradigm inspired SCARPS. The study also ignored model applicability.

Chughtai and Zayed (2008) predicted sewer pipeline deterioration using multiple regression model. The model included independent variables such as pipe material, depth, length, age, diameter, bedding, road type, and slope. The study selected key variables using the best subset analysis. F-test, t-test, residual analysis, lack of fit test, and Durbin-Watson test were used to determine variable significance. Four regression models predicted concrete, asbestos, cement, and PVC pipe conditions. The results revealed 72–88% accuracy and suggested inspecting pipes with extremely steep bed slopes first.

Tran et al. (2009) employed multivariate logistic regression to model pipe structural conditions. CCTV data from a Melbourne local government authority was used to compare model predictions. This model used pipe size, age, depth, slope, trees, hydraulic condition, road type, and soil type as independent variables. Logistic regression was calibrated using maximum likelihood calibration. Neural network models are better at modeling sewer pipeline structural deterioration.

Lubini and Fuamba (2011) constructed a sewer system deterioration logistic regression model. The model was based on pipe age, diameter, material, length, and slope in a Quebec City case study. Independent variables were assessed using the overall model test, the strength of association, the likelihood-ratio test, and the Wald Test. This study created a maintenance and operational planning deterioration curve. However, the logistic regression model was not tested.

Salman and Salem (2012) modeled wastewater collection line deterioration using ordinal regression, multinomial logistic regression, and binary logistic regression. Pipe size, length, slope, age, depth, material, sewer function, and road class calibrated the models. Five ordinal regressions were created, and a likelihood-ratio test was utilized to establish dependent-independent variable relationships. No ordinal regression model met odd assumptions. Developed multinomial logistic regression had 52% accuracy. Only binary logistic regression predicted sewage pipe condition with 66% accuracy. Different deterioration curves and equations from this study help explain network pipe behavior. Confusion matrix and real data verified logistic regression models. Pipe size, length, slope, age, material, and sewer type were significant predictors in binary logistic regression.

Kabir et al. (2018) applied Bayesian logistic regression to forecast sewer pipeline structure. Calgary's 12,728 sewer mains were used to model the wastewater network. Pipe age, material, diameter, length, slope, depth, rim elevation, and up-invert were used to create the model. This study used Bayesian model averaging to identify significant factors and logistic regression to predict sewer pipe condition. The independent variables were tested using P-test, Wald Test, likelihood-ratio test, and Durbin-Watson test. Good and bad sewer pipe conditions were identified. The model's performance was verified via the confusion matrix. Since pipe data were grouped by material, this model could not predict pipe condition.

2.6.3 Machine Learning (ML) Prediction Model for Sewers.

In 1943, Warren McCulloch and Walter Pitts implemented the first AI work based on knowledge of brain physiology and function, propositional logic, and Turing's theory of computation. Charniak and McDermott (1985) stated, “Artificial intelligence is "the study of mental faculties through computational models.”

According to Luger (2009), artificial intelligence can be decomposed into several categories as described 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 network and genetic algorithms
- AI and philosophy

Artificial intelligence models classify dependent variables from independent variables using data. Recent research has used neural networks and machine learning to model infrastructure deterioration.

2.6.3.1 Neural Network and Genetic Algorithms

Neural network and genetic algorithms mimic brain neurons (Luger, 2009). The human brain-inspired neural network and genetic algorithms. The power of these computing models depends on the structure of their nodes, which work like the brain (Koehn, 1994)

Fuzzy set theory and neural networks (NNs) were used to model the deterioration of infrastructure facilities among neural networks and genetic algorithm techniques (Tran, 2006).

Tran et al. (2007) generated an underground wastewater pipeline neural network deterioration model. This study calibrated the model using Markov Chain Monte Carlo simulation. The neural network was also compared to numerous discrimination analysis techniques for ranking. This model included pipe age, size, depth, slope, tree amount, road type, soil type, and wetness. The study found that Markov chain-calibrated neural networks outperform backpropagation-calibrated ones.

Khan et al. (2010) developed a structural condition prediction model to assess sewage pipe characteristics' importance and impact. This study evaluated pipe conditions using backpropagation and probabilistic neural networks. This model uses Pierrefonds, Quebec data. The model used pipe material, diameter, depth, bedding, length, and age. A neural network may prioritize sewage-leading inspection and rehabilitation, according to the models.

Sousa et al. (2014) evaluated the structural deterioration of sewer pipelines using logistic regression. As independent variables, the model included conduit material, diameter, length, age, depth, and slope. In addition to using support vector machines and artificial neural networks, the model of deterioration was created in this study. According to studies, logistic regression had the lowest modeling correlation. Due to model overlap, the authors state that this investigation cannot determine the optimal model.

Hawari et al. (2016) produced a simulation-based model for assessing the condition of wastewater pipelines using the fuzzy analytical network process (FANP). The FANP used a weighted scoring system to evaluate sewer pipes and weighed the factors influencing pipeline evaluation.

Gheytaspour et al. (2018) anticipate wastewater treatment facility oxygen consumption with a neural network. Due to treatment facilities' environmental impacts from the poor operation, process variable volatility, and linear analytic difficulties, artificial intelligence algorithms such as artificial neural networks have received attention. Regression analysis determined the input wastewater's biological oxygen demand, chemical oxygen demand, and pH. Error analysis chose the best neural network topology for prediction. The best multilayer perception network features the sigmoid tangent training function, one hidden layer in the input and output, 10 training nodes, and a 0.92 regression coefficient. Regression coefficients show that neural networks may predict wastewater treatment facility performance.

2.6.3.2 Machine Learning Algorithms

Machine learning is programming computers to learn from data. Machine learning, defined by Arthur Samuel in 1959, allows computers to learn without being programmed (Geron, 2017). Machine learning can be categorized into supervised and unsupervised categories. Supervised learning is used for most data analysis in condition prediction studies. The computer software or algorithm is trained to investigate historical data that contains the output or target variable in supervised learning. The training is used to estimate prediction for new or unrecorded data. In unsupervised learning, the goal variable is not included in training data.

According to Bishop (2016), Figure 2-4 illustrates three broad classifications of machine learning based on the basic principles of learning.

- Supervised learning models use input-output pairs as training data.
- Unsupervised learning uses input variables without output variables.
- Reinforcement learning is like unsupervised learning, the model does not provide output variables, and targets must be predicted by trial-and-error methodology.

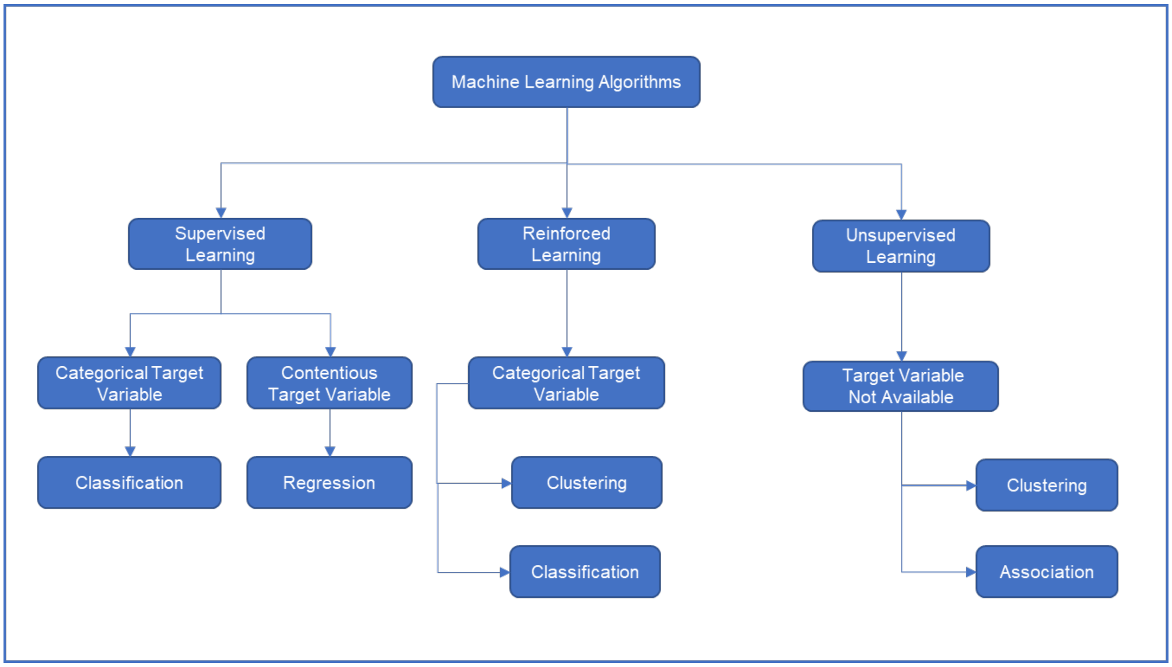

Machine Learning Algorithms
Supervised Learning
Reinforced Learning
Unsupervised Learning
Categorical Target Variable
Contentious Target Variable
Categorical Target Variable
Target Variable Not Available
Classification
Regression
Clustering
Classification
Clustering
Association

Figure 2-4 Prediction Model Classification. (Liu et al., 2022)

Based on the above-listed categories, the modeling can be further classified as below:

- Classification: supervised learning models the outputs of two or more classes.
- Regression: supervised, continuous outputs.
- Clustering classifies inputs into groups. Unsupervised, unlike classification and regression.

Machine learning is gaining popularity across numerous industries. In the wastewater business, machine learning models such as support vector machines (SVM), decision trees, random forests, and Bayesian regressions have been used to forecast sewage network damage.

Najafi and Kulandaivel (2005) created an ANN-based condition prediction model in 2005. To train the model, age, length, size, material type, depth, slope, and sewer type were independent variables. The model performed well during training but poorly during testing. The study noted that further statistical analysis was needed with more significant data.

Probabilistic neural networks modeled stormwater pipe structural deterioration by Tran et al. (2007). The model employs 650 data points from Greater Dandenong, Victoria, Australia. This model incorporated pipe diameter, age, depth, slope, location, number of trees, hydraulic condition, soil type, and wetness. Probabilistic neural networks predicted pipeline deterioration better than discriminant approaches.

Stormwater pipe prediction was studied by Tran et al. (2009). Comparing model predictions with Melbourne local government CCTV data. The independent factors were pipe size, age, depth, slope, trees, hydraulic condition, road type, and soil type. Maximum likelihood and neural network calibrated logistic regression model. Neural networks model wastewater pipe structural deterioration better.

Mashford et al. (2011) predicted sewage pipeline grade with a support vector machine. CCTV footage from South Australia’s Adelaide wastewater collection network established the model's predictive performance. The sewer pipe condition was scored 1–5 (excellent–bad). Pipe diameter, age, road type, slope, start/end invert, material, soil type, soil corrosivity, grade, angle, sulfate soil, and groundwater level were input factors. The support vector machine's 91 percent prediction accuracy makes it a promising tool for sewage pipe damage simulation. The study lacked sufficient condition data, according to the authors.

Harvey and Mcbean (2014a) employed random forests to forecast sanitary sewage pipe structural status. The sewer database came from Guelph, Ontario, Canada. Pipe age, material, length, diameter, service type, slope, up elevation, down elevation, depth, land use, and road type were used to build the model. The research showed that random forest models can accurately predict sewer pipe conditions with an area under the ROC curve of 0.81. The cost and time of projects can be reduced by using random forest prediction models to estimate the status of uninspected sewer systems.

Another article by Harvey and Mcbean (2014b) used support vector machine and decision tree models to schedule sewer pipeline inspections. Data from Guelph, Ontario, Canada, was utilized to create the model. Pipe material, age, kind of sewer, diameter, length, slope, down elevation, depth, and road coverage were the model's inputs. Recent results have indicated 76% of the time the support vector machine accurately predicted wastewater pipe condition ratings. Nevertheless, decision trees were beneficial for prioritization and organizing inspections of wastewater pipe systems.

Hernandez et al. (2017) created a structural condition prediction model by combining logistic regression, random forests, multinomial logistic regression, linear discriminant analysis, and support vector machine. The study compared model performance. The sole performance measures were true positive and false positive rates. Prediction and planning of sewer pipe inspections were also bad.

Laakso et al. (2018) estimated sewer pipeline conditions using random forest and binary logistic regression. This study investigated pipe degradation. This study used southern Finnish databases. EN-13508-2 evaluated sewer pipes. Score 0 denoted "no defect" and 4 "serious defect". The model considered pipe age, diameter, material, slope, depth, length, soil type, road class, distance to tree, intersection with stormwater or water supply pipes, and yearly sewage discharge. Binary logistic regression scored 62%, and random forest 67%. Logistic regression and random forest models predicted sewer pipeline conditions.

Malek Mohammadi et al. (2019) used decision trees, random forests, and gradient-boosting trees to construct sanitary sewage pipe condition prediction models 2020. Gradient boosting tree-based model accuracy was 87%. The model classified pipe condition ratings as binary rather than multi-class, yet accuracy was good. The study suggested future research on multi-class condition prediction, which would benefit the municipality during inspection and condition assessment.

Loganathan (2021) developed a sanitary sewer pipe condition prediction model using Logistic Regression (LR), k-Nearest Neighbors (k-NN), and Random Forest (RF), which are supervised machine learning algorithms. His results showed that RF outperformed LR and k-NN. Pipe variables like Soil and slope are ignored. The prediction model used sewer pipe data from one city.

Atambo (2021) built a prediction model for sewer pipes by utilizing Multiple Logistic Regression (MLR) and Artificial Neural Networks (ANN) for the inspection data collected for one city with 2616 datasets. The pipe variable considered in this study includes- Pipe age, material, diameter, length depth, and Slope. The prediction model showed that ANN outperformed the MLR accuracy. The research recommended collecting sewer pipe inspection data from different cities to make it a comprehensive prediction model that the city municipalities can employ for prioritizing the inspection for the future.

Shirkhanloo (2022) constructed the prediction model for sewer pipes by using the inspection sewer data for the City of Dallas. Supervised learning algorithms such as Bagging and Boosting algorithms were employed in creating a model. The outcome showed that Random Forest (RF) outperformed the other algorithms. The pipe variables like – age, material, length, soil, and slope were considered. This research collected inspection data for just one city, and this research cannot be employed as a tool to prioritize the inspection of sewer pipes.

In recent years, a number of models for predicting the condition of a sewer system have been devised and discussed in this chapter. The summary of the developed prediction models to date can be seen as shown in Table 2-2.

Table 2-2 Summary of Developed Sewer Condition Prediction Models

	Authors	Year	Model	Variables included	Condition Assessment Standard	Condition Rating Output	Number of Data
1	Ariaratnam et al.	2001	• Logistic regression	Age, Material, Diameter, Depth,	1,2,3,4,5	WRc	748
2	Najafi and Kulandaivel	2005	• Neural network	Age, Material, Diameter, Depth, Length, Slope	1,2,3,4,5	PACP	-
2	Tran et al.	2006	• Neural network	Age, Diameter, Depth, Length, Slope	1,2,3	WSAA	583
3	Tran et al.	2007	• Neural network • Multiple discrimination analysis	Age, Diameter, Depth, Length, Slope	1,2,3	WSAA	150
4	Chughtai and Zayed	2008	• Linear regression	Age, Material, Diameter, Depth, Length, Slope	1,2,3,4,5	WRc	-
5	Tran et al.	2009	• Neural network • Ordered probit model	Age, Diameter, Depth, Length, Slope	1,2,3	WSAA	417
6	Khan et al.	2010	• Neural network	Age, Diameter, Depth, Length, Slope	1,2,3,4,5	WRc	200
7	Lubini and Fuamba	2011	• Logistic regression	Age, Diameter, Depth, Length, Slope	1,2,3	PACP	459
	Mashford et al.	2011	• Support vector machine (SVM)	Age, Material, Diameter, Slope	1, 2, 3	PACP	1,441
8	Salman and Salem	2012	• Ordinal regression • Logistic regression • Binary regression	Age, Diameter, Depth, Length, Slope	1,2,3,4,5	PACP	11,373
9	Syachrani et al.	2013	• Decision tree • Neural Network	Age, Material, Diameter, Length, Slope	1,2,3,4,5	PACP	52,855
10	Sousa et al.	2014	• Neural network • Support vector machine • Logistic regression	Age, Material, Diameter, Depth, Length, Slope	1,2,3,4,5	PACP	745
11	Harvey and McBean	2014(a)	• Random forest • Decision Tree • Support vector machine	Age, Material, Diameter, Depth, Length, Slope	1,2,3,4,5	WRc	1,825
12	Bakry et al.	2016	• Multiple regression	Age, Material, Diameter, Depth, Length, Slope	1,2,3,4,5	PACP	84
13	Gedam et al.	2016	• Linear regression	Age, Material, Diameter, Depth, Length, Slope	1,2,3,4,5	PACP	155
14	Hernandez et al.	2017	• Logistic regression • Random forest	Age, Material, Diameter, Depth, Length, Slope	1,2,3,4,5	PACP	4,327
15	Kabir et al.	2018	• Bayesian logistic regression	Age, Material, Diameter, Depth, Length, Slope	1,2,3,4,5	PACP	12,728
16	Laakso et al.	2018	• Binary logistic regression • Random forest	Age, Material, Diameter, Depth, Length, Slope	1, 2, 3, 4	EN-13508-2	6,700
17	Malek Mohammadi	2019	• Logistic regression • k-NN • XGBoost	Age, Material, Diameter, Length	1,2,3,4,5	PACP	20,282
18	Mazumder et al.	2020	• k-NN • Decision tree • Random forest	Age, Material, Diameter, Depth, Length, Slope	1,2,3,4,5	PACP	959

			• AdaBoost, XGBoost, LGBoost, CATBoost				
19	Loganathan	2021	• Logistic regression • k-NN • Random forest	Age, Material, Diameter, Depth, Length, Slope	1,2,3,4,5	PACP	32,751
20	Shirkhanloo	2022	• Logistic regression • Decision tree • Random forest • AdaBoost • XGBoost • Gradient Boost	Age, Material, Diameter, Depth, Length, Slope	1,2,3,4,5	PACP	3,376

2.7 Chapter Summary

As previously indicated, pipe deterioration is complex and cannot be caused by a single factor. In addition, wastewater organizations and municipalities typically lack the financial resources to regularly inspect all network pipelines. To reduce inspection costs and provide a comprehensive plan for prioritization and inspection scheduling, an alternative strategy must be adopted. This chapter describes numerous wastewater pipe deterioration models and factors. However, individual sewer pipe condition prediction models have not been exhaustively investigated, and most research indicates that novel data analysis methods can predict the behavior and condition of sewer pipelines. The objective of this dissertation is to develop a comprehensive sewer pipe prediction model to identify the critical factors responsible for their deterioration.

CHAPTER 3

PREDICTION MODEL DEVELOPMENT

3.1 Introduction

Chapter 3 discusses building a condition prediction model through different machine-learning algorithms. (Mohri et al., 2018) described Machine learning as a broad term for computational algorithms that rely on prior knowledge to generate accurate predictions. In this instance, previous understanding relates to the stored historical data from which a computer program can learn, a process known as algorithm training. As discussed in previous chapters, the study implements supervised learning techniques to construct a prediction model.

Regression and classification are the main supervised machine learning methods. Regression occurs when many independent factors must predict a continuous dependent variable (Müller and Guido, 2016). The dependent or outcome variable in this study is categorical with five classifications. This study develops a model using classification machine learning techniques. Models are trained using processed data, as discussed in the previous chapter. Python, a prominent data science programming language, is utilized to create prediction models in this study. Python's open-source nature and many free add-on packages make it appealing.

The number and type of independent variables, as well as the dependent variable, affect the sewer conduit prediction models. It is crucial to select a predictive model designed for predicting dependent factors with multiple classes, given that the dependent variable in this study is the condition rating of sewer pipelines, which has been organized into five categories. Therefore, the best models for this research were selected based on their ability to predict multi-categorical dependent variables.

Using Python software, this dissertation develops a statistical model utilizing logistic regression and classification. For analyzing datasets with two or more discrete outcome variables, the classification model is the most widely used. decision tree and random forest models with default parameters were the first set of models developed in this research to evaluate their performance score (accuracy and F1-score).

The second set of models devised in this study are hyperparameter tuning of decision trees and random forest algorithm to improve the performance. Python was chosen for this study because it is an open-source programming language with various free add-on libraries. As a third type of model, tree-based boosting models are created. They are among the most effective learning techniques presented and are intended for classification problems. In this dissertation, decision Tree, bagging techniques, such as random forest, and boosting techniques, such as AdaBoost, Gradient Boosting Tree, and XGBoost, are developed and explained.

3.2 Model Selection

Model selection is crucial to statistical analysis since numerous factors affect regression models. Sewer pipe deterioration models depend on information data, independent variables, and dependent variables. As mentioned, sewer pipe values classify condition prediction scales. Choose a predictive model that can forecast categorical dependent variables.

This research study aims to forecast the future condition states of individual sewer pipelines. It is seen that the condition states of sewer pipes are typically described as discrete or categorical values which are not serial numbers and are classified with 5 different classes; consequently, linear and exponential regressions are not appropriate for predicting categorical variables because they minimize the total number of squared distances between the predicted and actual condition ratings (Salman 2010). So, the classification type of regression techniques is employed to construct the model.

In this dissertation, the most ideal models are selected on the below-listed criteria:

- Model performance in predicting categorical dependent variables.
- The model's capability to be trained by nominal variables.
- The model's end results.

3.3 Correction of Imbalanced data

In EDA of Chapter 4, it will be presented that the PACP score of 4 and 5 has its rare occurrence when compared to the PACP score of 1, 2 and 3. This variance in the dataset is termed an imbalanced state. When imbalanced data is used in traditional classification algorithms, it can often lead to poor model performance (Teh et al., 2020). In general, the minority class must be accorded a higher priority when dealing with unbalanced datasets, as the repercussions of misclassifying a minority class are exponentially more significant than those of the other classes. In this study, a PACP score of 4 and 5 are given more weight because interpreting it as a score of 1. 2 and 3 would be more detrimental.

Data scientists and researchers explored different techniques to treat this imbalance in dataset and found that classification algorithms like logistic regression, support vector machine, and decision tree can be successfully used in training the dataset (Hosmer et al., 2013). This study employs the data resampling method to treat the imbalance datasets which is the most effective in replicating or removing the data points to make the majority class and minority class meet the requirement. The data resampling technique is divided into (1) random under-sampling and (2) random over-sampling.

3.3.1 Random Under-samplings

In this technique, the data points/observations are removed from the majority class randomly to match the minority class. It should be noted in this procedure that the dataset may lose essential information during the removal process.

3.3.2 Random Over-samplings

In this technique, the data points/observations are randomly replicated in the minority class to match the majority class. This method is incredibly useful when the dataset has low minority class observations. It should be noted that over-sampling may result in overfitting models. Figure 3.1 depicts the random under-sampling and random over-sampling concepts.

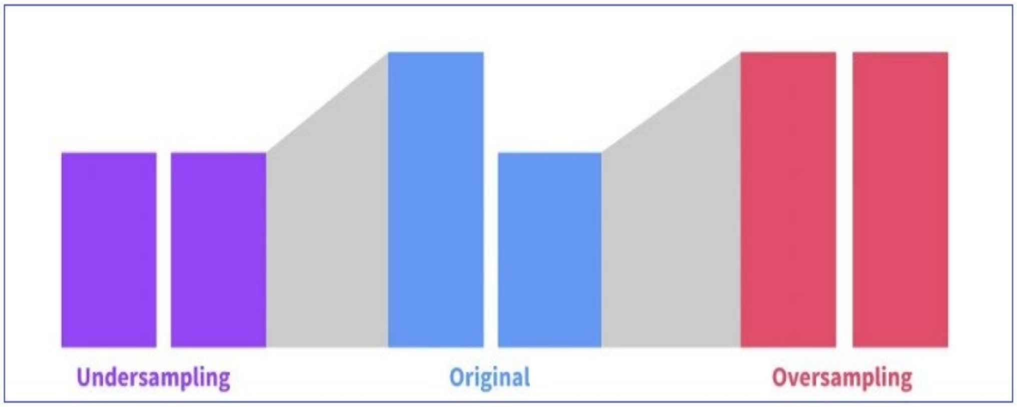


Figure 3-1 Random Under-Sampling and Random Over-Sampling Concept

3.4. Hyper-parameters Tuning

The process of finding optimal hyperparameters for a model is known as hyperparameter Tuned. Hyperparameters are the parameters that govern the entire training process. The hyperparameter values are set at the beginning of the learning process begins. Selecting optimal hyperparameters can lead to increases in the overall model's performance and can help in reducing both overfitting and underfitting and will have a substantial effect on the model's performance.

Finding the optimal set of hyper parameter values for models with many hyperparameters can be a time-consuming endeavor. To make the procedure more efficient, two of the most prevalent methods are available in sklearn: Grid-Search and Random-Search. Table 3-1 presents the built in hyperparameters in python for tree-based models developed in this study. While building the tree-based models, the below mentioned hyperparameters are set to get better performance.

Table 3-1 Built in Hyperparameters for Tree-based Models

Decision Trees	**Random Forest**	**Adaboost**	**Gradient Boost**	**XGBoost**
max_depth	n_estimators	base_estimator	Learning_rate	Learning_rate
min_samples_split	max_features	n_estimators	gamma	gamma
min_samples_leaf	max_depth	learning_rate	scale_pos_weight	scale_pos_weight
max_features	min_samples_split	algorithm	N_estimators	colsample_bytree
class_weight	min_samples_leaf	classes	max_depth	colsample_bylevel
	bootstrap	estimator_weights	min_samples_split	colsample_bynode
		estimator_errors	max_features	max_features

3.4.1 Grid-Search

Grid-Search is a technique for finding the optimal set of hyperparameters for a model from a search space. It iterates over all black circles in a sequence, determining the best set based on the best score obtained. Grid-Search is optimal when the search space is limited, and it can be used when there are no time constraints and obtain finest results (Liashchynskyi, P. and Liashchynskyi, P. 2012). Figure 3-2 shows the Grid-Search space illustration. In the Grid-search method of cross validation, the parameters range are manually set, and the algorithms makes a complete search over the data set.

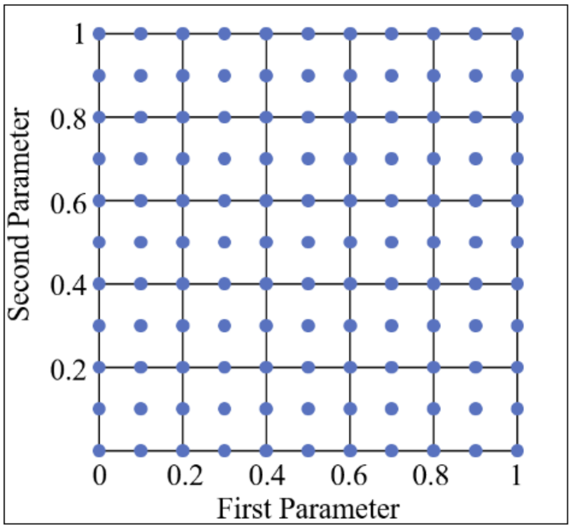


Figure 3-2 Illustration on Grid-Search Space
(Liashchynskyi, P. and Liashchynskyi, P. 2012)

3.4.2 Random-Search

Random-Search replaces the exhaustive selection of all combinations applied readily to discrete cases by generalizing to continuous and mixed spaces. Random-Search can outperform Grid-Search, particularly if a limited number of hyperparameters influence the performance of the machine learning algorithm. Random-Search is optimal when the search space is large. Randomized Search is known to produce superior results to Gridsearch. Figure 3-3 illustrates Random-Search Space.

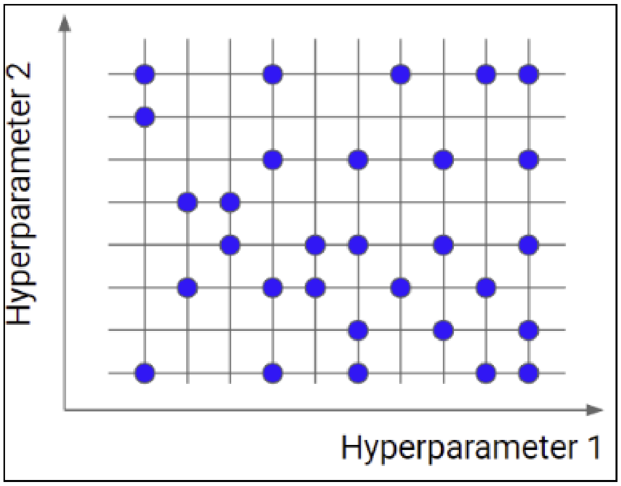


Figure 3-3 Illustration on Random-Search Space
(Liashchynskyi, P. and Liashchynskyi, P. 2012)

3.5 Prediction Model Algorithms

3.5.1 Decision Tree

Decision trees are a rule-based method for addressing classification and regression issues. Using the values in each feature, they divide the dataset so that all data points with the same class are grouped together. Key terms in decision tree are listed below.

- 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., 2017). Gini index is calculated by equation 3.15 (Geron, 2017).

$Gi\ (n) = 1 - \sum_{j=1}^{2}\left(p_j{}^2\right)$………………………………………………………….. Eq. 3.1

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 sprout from it represent answers. Leaf nodes sprout from decision nodes in the same way as a leaf. sprouts from a tree branch. Figure 3-4 shows the elements of a decision tree.

The DT is composed of a root node which is the topmost node acting as a parent node to branch nodes (subtree), and leaf nodes. On each internal node, an attribute is tested; the test result is displayed on the branch, and the class label is displayed on the leaf node. A decision tree is a tree in which each node represents a feature (attribute), each branch represents a decision (rule), and each leaf represents a result (categorical or continuous value) (Patel, H., and Prajapati, P. (2018)).

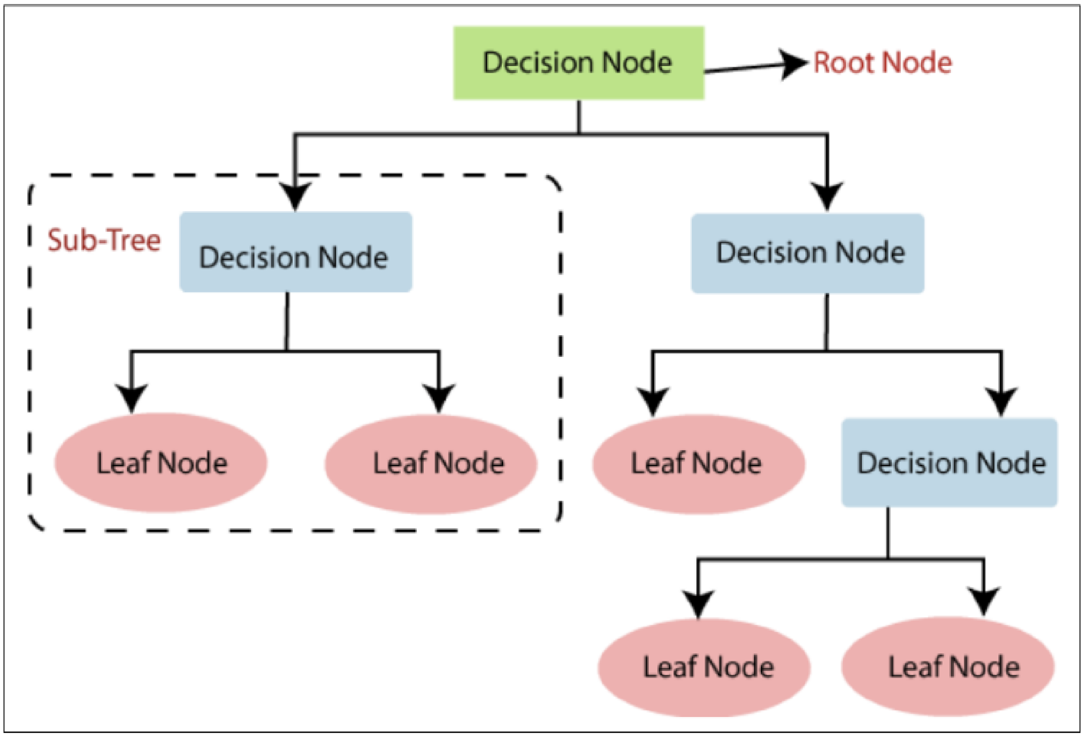


Figure 3-4 Decision Tree Classification Algorithm
Patel, H., and Prajapati, P. (2018)

3.5.2 Bagging Technique-Random Forest

Even though Decision Tree (DT) is regarded as an effective supervised learning algorithm for classification problems, one of its most common limitations is that it tends to overfit the training data (Müller and Guido,2016). To surmount DT's limitations, the RF method is employed in this study. When creating the RF tree, a dataset should be supplied to the training set. In DT, even if each tree performs a good job of prediction, it will undoubtedly overfit some portion of the data. The level of overfitting could then be constrained by aggregating the results of numerous trees that perform well and overfit in several ways (Rokach, L., and Maimon, O. (2008). The final aggregation of numerous DTs with retained predictive ability is called RF. Figure 3-5 presents the operational structure of Random Forest classification.

Random forests evaluate divisions in decision trees based on the Gini index. It efficiently indicates impurity, and how well a split separates different classes in index are computationally efficient, which makes it appropriate for simultaneously constructing multiple decision trees. It is less sensitive to imbalanced class distributions, making it helpful in handling datasets in which one class predominates.

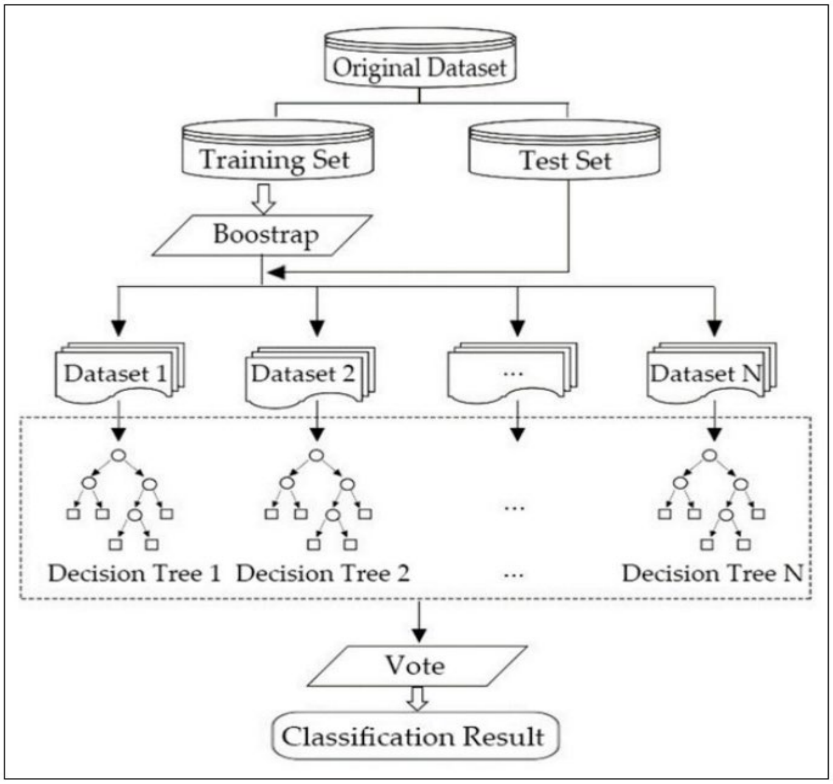


Figure 3-5 Operational structure of Random Forest classification Algorithm

Overfitting the training data is one of the decision tree's most prevalent limitations. Overfitting is a statistical modeling error that happens when a function is overly compatible with a limited set of data points. Therefore, the model only applies to its initial data set and not to other data sets (Müller and Guido, 2016). Pre-pruning is a common strategy for preventing overfitting in decision trees. It entails limiting the tree's maximum depth and number of leaves, but it is not always a solution for a decision tree that has been overfitted. To solve this issue, the random forest method is suggested.

3.5.3 Tree-Based Models-Boosting Algorithm

"Boosting" is a general technique for enhancing the efficacy of any learning algorithm. Theoretically, boosting can be used to substantially reduce the error of any "weak" algorithm for learning that reliably generates classifiers that are only marginally superior to random guessing. Boosting's actual practical value can only be determined by testing the method on existing machine learning problems, despite the theoretical results' predictions of its prospective advantages. Boosting operates by repeatedly applying a given weak learning algorithm to different distributions over the training data and then aggregating the classifiers produced by the weak learner into a single composite classifier (Ross Quinlan, J. 1996).

The boosting algorithm is also known as the Meta algorithm. The accuracy of the boosting algorithm is often found to be overperforming compared to Random Forest bagging algorithms. The primary difference between the building of bagged trees and boosted trees is that we now replace the (random) sampling with some form of *weighting* in which instances are assigned weights, and the weights of the *nth* tree are dependent on the outputs returned by the previously created (nth-1) tree model (Hastie et al. 2017).

3.5.3.1 AdaBoost

Freund and Schapiro (1999) devised the most prominent boosting algorithm known as AdaBoost (adaptive boosting). Figure 3-6 illustrates the visual representation of the enhancing AdaBoost algorithm`

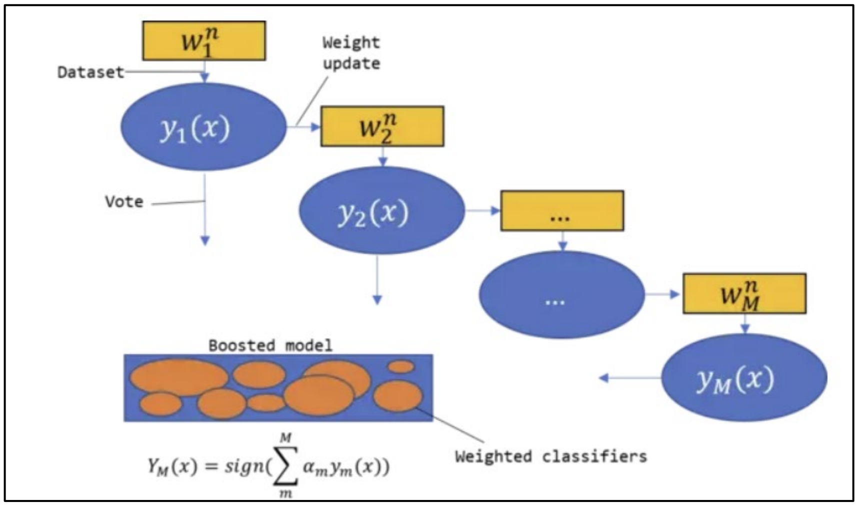


Figure 3-6 Visual Representation of Enhancing AdaBoost Algorithm
(Freund and Schapiro, 1999)

Here, the various base classifiers are each built on a weighted dataset, where the weights of the individual instances in the dataset depend on the results the preceding base classifiers had produced for these instances. Suppose if the instance is incorrectly classified, then the weight for this instance will be enhanced in the subsequent model. In contrast, the weight will remain unchanged if the classification is accurate. The final decision is made through a weighted majority of the base classifiers, with the weights based on the inaccurate classification rates of the models. A model with an excellent ability to classification will receive a high weight, while one with a low classification accuracy will receive a low weight. Boosting pseudocode is explained below.

Initialize all weights to w=1/n where “n” is the number of instances in the dataset.

Consider t < T (T== total models to be built)

Build a model to design a hypothesis h_t (x_n) for the x_n data points in the training set.

Error calculation ∈, is calculated as equation 3-2

$$\in_t = \frac{\sum_{n-1}^{N} w_n^t * I\,(y_n \neq h_t(x_n))}{\sum_{n-1}^{N} w_n^{(t)}} \qquad \text{Eq 3-2}$$

where I (Cond) returns 1 if I (Cond) == True and 0 otherwise

Compute α with equation 3-3,

$$\alpha_t = \log \frac{(1-\epsilon t)}{\epsilon_t} \qquad \text{Eq 3.3}$$

Update the weights for the N training instances in the next (t+1) model with equation 3-4:

$$w_n^{(t+1)} = W_{n^*}^t \exp\left(\alpha_{t^*} - \perp\,\left(y_n \neq h_t(x_n)\right)\right) \qquad \text{Eq 3-4}$$

The final output is calculated after the T iterations by equation 3-5,

$$f(x) = \operatorname{sign}\left(\sum_t^T \alpha_t * h_t(\varkappa)\right) \qquad \text{Eq 3-5}$$

3.5.3.2 Gradient Boost Algorithm

Gradient boosting is a method that stands out for its prediction speed and precision, especially with large and complex datasets. This algorithm's central concept is to sequentially construct models, with each successive model attempting to reduce the defects of its predecessor. Gradient Boosting Regressor is utilized when the target column is continuous, whereas Gradient Boosting Classifier is utilized for classification problems. The "Loss function" is the only distinction between the two. The aim here is to limit this loss function using gradient descent and weak learners. Since it is founded on the loss function, there will be different loss functions for regression problems, such as Mean squared error (MSE), and classification problems, such as log-likelihood (Geron 2017).

Gradient boosting begins with the construction of a base model to predict the observations in the training dataset. In this step, the average if the target model/dependent values are calculated to set the base model and mathematically, it is expressed in Equation 3-6 below (Malek Mohammadi, 2019):

$$F_0(x) = \arg_{\gamma} \min \sum_{i=1}^{n} L(y_i, \gamma) \qquad \text{Eq 3-6}$$

Where, L, is our loss function.

$_{\gamma}$ is the predicted value,

and arg min means calculating the predicted value/gamma for which the loss function is minimum.

Finally, the loss function presented in the below equation 3-7,

$$L = \frac{1}{n} \sum_{i=0}^{n} (y_i - \gamma_i)^2 \qquad \text{Eq 3-7}$$

Here y_i is the observed value,

And the minimum value of $_{\gamma}$ such that this loss function is minimum and shown in equation 3-8,

$$\frac{dL}{d\gamma} = \frac{2}{2} \left(\sum_{i=0}^{n} (y_i - \gamma_i) \right) = -\sum_{i=0}^{n} (y_i - \gamma_i) \qquad \text{Eq 3-8}$$

3.5.3.3 XGBoost Algorithm

XGBoost is a popular gradient-boosting implementation with unique features such as regularization, handling sparse data, weighted quantile sketch, block structure for parallel learning, cache awareness, and out-of-core computing. It penalizes complex models through L1 and L2 regularization, handles sparse data, and optimizes disk space for large datasets. XGBoost's distributed weighted quantile sketch algorithm effectively handles weighted data, allowing for faster computing on multiple CPU cores. Additionally, XGBoost optimizes hardware usage by allocating internal buffers in each thread for gradient statistics. XG Boost's efficient handling of missing values is one of its essential features, allowing it to handle data from the real world with missing values without requiring extensive preprocessing. In addition, XGBoost has built-in support for parallel processing, allowing large datasets to be trained in a reasonable period.

3.6 Model Evaluation

3.6.1 Evaluation Criteria

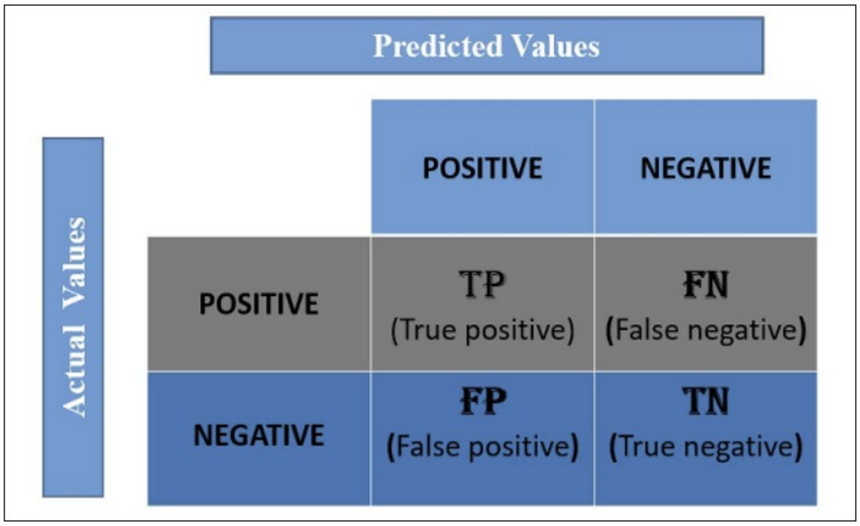
The objective of supervised learning techniques, which are frequently trained by a set of data, is to construct a predictive model. Prediction model performance must always be evaluated to assess the model's level of accuracy and define its key parameters. There are numerous ways to evaluate the efficacy of machine learning models. Using a confusion matrix, Receiver Operating Characteristic (ROC), and Area Under the Curve (AUC), this research evaluates tree-based models.

3.6.2 Evaluation Metrics

In this section of the dissertation, various evaluation methods are examined in depth. The type of output anticipated from the classification model would guide the selection of a particular metric.

3.6.2.1 Confusion Matrix

The confusion matrix is a tabular representation of your prediction model's performance. Every value in a confusion matrix represents the proportion of predictions where the model correctly or incorrectly classified the classes. It is applicable to both binary classification and multiclass classification problems. Figure 3-7 represents the example of binary confusion matrix.



		Predicted Values	
		POSITIVE	NEGATIVE
Actual Values	POSITIVE	TP (True positive)	FN (False negative)
	NEGATIVE	FP (False positive)	TN (True negative)

Figure 3-7 Binary Class Confusion Matrix (Malek Mohammadi (2019)

The confusion matrix is employed to optimize machine learning models. The confusion matrix is a N x N matrix, with N representing the number of classes or outputs. For two classes, a two-by-two disorientation matrix is developed and for three classes, a 3 x 3 matrix is generated. Confusion matrices represent the number of actual and predicted values. The output "TN" indicates the number of negative examples that were accurately classified as negative. Similarly, "TP" stands for True Positive and represents the total number of positive instances that have been correctly classified. The term "FP" indicates a False Positive value, indicating the number of actual negative examples misclassified as positive, whereas "FN" indicates a False Negative value, which is the number of actual instances that are positive misclassified as negative. Accuracy is one of the most frequently employed metrics when conducting classification. The precision of a model (as measured by a confusion matrix) is computed using the following formula (Kulkarni et al. 2020).

3.7.2.2 Accuracy

Accuracy can be deceiving when applied to unbalanced datasets; therefore, other metrics based on the confusion matrix can be utilized to evaluate performance. The "confusion matrix ()" function of the "sklearn" library in Python can be utilized to obtain the confusion matrix. This function may be imported into Python with the command "from SKlearn. metrics import confusion matrix." To obtain the confusion matrix, users must provide the function with both actual and predicted values (Kulkarni, A., et al. 2020).

3.7.2.3 Precision and Recall

Precision and recall are extensively employed and popular classification metrics. Precision indicates how well the model predicts positive values. Thus, it gauges the accuracy of a positive outcome prediction. Also referred to as the positive predictive value. Recall is beneficial for measuring a model's ability to predict positive outcomes, and it is also known as a model's sensitivity. Both measures provide valuable information, but the goal is to increase recall without compromising precision. Using the "precision score ()" and “recall score ()" functions, respectively, precision and recall values can be calculated in Python. The formulas for evaluating precision and recall are given below equation 3-9 and 3-10:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{Eq 3-9}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{Eq 3-10}$$

3.7.2.4 F1-Score Binary Class

F1-score is derived by calculating the harmonic mean of precision and recall and combining them into a singular metric to evaluate the performance of the classification model. The F1-score is measured on a scale from 0 to 1, with 1 denoting superior model performance and 0 denoting inferior model performance. Since the F1-score is a weighted average of both precision and recall, both factors contribute equally. Consequently, it can be used to determine the optimal method to compromise between the two values. According to Hossin and Sulaiman (2015), F1-score is a crucial metric for determining the efficacy of a developed model based on evaluation criteria. As shown in Equation 3-11, the

$$\text{F1} = \frac{2(\text{R})(\text{P})}{(\text{R}+\text{P})} \qquad \text{Eq. 3-11}$$

3.7.2.5 F1-Score Multi Class

F1-Score should include all classes in instances involving multiple classes. To accomplish this, a multi-class Precision and Recall measure must be incorporated into the harmonic mean. These metrics may have two distinct specifications, resulting in two distinct metrics: The micro and macro F1-scores (Grandini et al. 2020).

The Micro and Macro F1- scores are computed as presented in the below equations from Eq. 3-12 through Eq. 3-17 below:

$$\text{Micro Average Precision} = \frac{\sum_{k=1}^{K} \text{TP}_k}{\sum_{k=1}^{K} \text{Total Column}_k} = \frac{\sum_{k=1}^{K} \text{TP}_k}{\text{Grand Total}} \qquad \text{Eq. 3-12}$$

$$\text{Micro Average Recall} = \frac{\sum_{k=1}^{K} \text{TP}_k}{\sum_{k=1}^{K} \text{Total Row}_k} = \frac{\sum_{k=1}^{K} \text{TP}_k}{\text{Grand Total}} \qquad \text{Eq. 3-13}$$

$$\text{Micro Average F1-Score} = \frac{\sum_{k=1}^{K} \text{TP}_k}{\text{Grand Total}} \qquad \text{Eq. 3-14}$$

$$\text{Macro Average Precision} = \frac{\sum_{k=1}^{K} \text{Precision}_k}{\text{K}} \qquad \text{Eq. 3-15}$$

$$\text{Macro Average Recall} = \frac{\sum_{k=1}^{K} \text{Recall}_k}{\sum_{k=1}^{K} \text{K}} \qquad \text{Eq. 3-16}$$

$$\text{Macro F1-Score} = 2 * \left(\frac{\text{MacroAverage Precision} * \text{MacroAverage Recall}}{\text{MacroAverage Precision}^{-1} * \text{MacroAverage Recall}^{-1}}\right) \qquad \text{Eq. 3-17}$$

3.7.2.6 Receiver Operating Characteristic (ROC) and Area Under the Curve (AUC)

The ROC (receiver operating characteristic) curve is employed to assess a classifier's performance. On the x-axis, the false positive rate (FPR) is plotted against the true positive rate (TPR) on the y-axis. For a classifier to execute classification, a range of thresholds from 0 to 1 is defined. FPR and TPR are plotted against one another for every point. An ROC curve illustration is shown in Figure 3-8. The upper left corner of the ROC curve indicates decent classification, while the lower right corner indicates poor classification. A classifier is deemed effective if it reaches the upper left corner. The diagonal in the graph represents guessing at random. If the ROC curve of a classifier is below the diagonal, that classifier performs worse than random guesswork, which completely defeats the purpose. Consequently, it is anticipated that the ROC curve will always be in the upper diagonal. A ROC curve is useful because it provides a graphical representation of a classifier, but it is always advisable to compute a numerical score for a classifier. Calculating the area under the curve (AUC) for the ROC curve is common practice. The AUC value represents a score ranging from 0 to 1. Any classifier whose ROC curve lies below the diagonal will receive an AUC score below 0.5. Similarly, ROC curves in the upper diagonal will have AUC values greater than 0.5. A perfect classifier will have an AUC score of 1, which corresponds to the upper-left corner of the diagram.

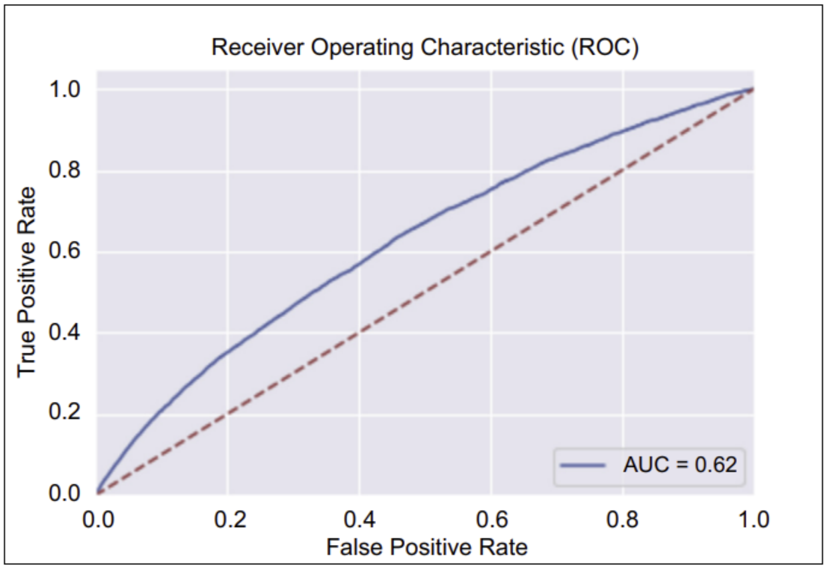


Figure 3-8 An Example of ROC and AUC
(Shirkhanloo, 2022)

3.7 Feature Importance in Tree-Based Models

Several useful properties can be extracted to summarize the operation of the tree. The most popular is feature importance, which determines the significance of each feature for the decision a tree makes. Each variable is represented by a number between 0 and 1, with 1 indicating that the variable completely predicts the objective and 0 indicating that it is not used at all. The weight values of features always sum up to 1 (Geron, 2017).

According to Biau and Scornet (2016), two measures of significance, Mean Decrease Impurity (MDI) and Mean Decrease Accuracy (MDA), are typically employed to evaluate the significance of the variables. MDI keeps track of the number of times an independent variable is utilized to divide a node. Using the Mean

Decrease Impurity (MDI), the weighted decrease of impurity from splitting on the variable, averaged across all trees, is measured. Mean Decrease Accuracy (MDA), on the other hand, is determined by aggregating the variation in out-of-bag error estimation between before and after the permutation across all trees.

3.7 Chapter Summary

This chapter discussed the treatments for imbalanced datasets, such as resampling techniques, to prepare data for training with machine learning algorithms. As discussed in subsequent chapters, under sampled and oversampled datasets will be used to train and develop condition prediction models. Various supervised learning algorithms, including logistic regression – decision tree, bagging algorithm like Random Forest and boosting algorithms like Gradient Boost, AdaBoost and XGBoost used in this investigation are also discussed.

Based on the confusion matrix developed for all these above-mentioned methods, it was found that the models trained with imbalanced dataset failed to classify structurally poor condition pipes. It was found that all three methods performed better than imbalanced when trained with over- sampled dataset.

CHAPTER 4

DATA COLLECTION AND PREPARATION

4.1 Introduction

The asset management system of sanitary sewers involves two essential major activities such as pipe inspection and condition evaluation, to replace/rehabilitate to make them fully functional to serve their purpose. In recent years, city municipalities in the United States have employed CCTVs to perform the sanitary sewer pipe inspection (NASSCO, 2015). This study's scope is restricted to gravity-flow sanitary sewer pipelines, excluding force main systems. The condition of pipes is assessed by utilizing the guidelines given out by Pipeline Assessment and Certification Program (PACP). The PACP assessment method has scored the pipe condition on a scale from 1 to 5, with 5 indicating a pipe is failing condition and 1 with no defects. CCTV inspection data included a breakdown of the main inspection and a score for each individual pipe segment based on the PACP condition rating system. For both structural and operational conditions, information such as pipe rating, fast rating, and pipe rating index was available for each pipe segment. Additionally, this database contained information regarding the general condition of pipelines. Geographic information system (GIS) databases are utilized to maintain the sewer system inventory. The recorded database inventory contains details regarding pipe installation, pipe location about, geographical maps, etc. The sewer system inventory is managed using Geographic Information System (GIS) databases. The recorded database inventory includes information about pipe installation, pipe location in relation to geographical maps, etc. (Shirkhanloo, 2022).

This section of the dissertation presents the data collection, data preparation for using them in the prediction model algorithms and showing the results from Exploratory Data Analysis (EDA). The EDA generates histograms displaying the frequency of variables which are used to compare the factors that influence the condition of sewer conduit. EDA also presents descriptive statistics and correlation analysis of the data. The study involves collecting sewer pipe inspection data from 11 city municipalities all over the United States. The primary purpose of this research is to collect the sewer pipe data from different utilities and compare them to have diversified data to construct a comprehensive prediction model which the utility owners and cities can practically implement to prioritize the sewer pipe inspection activity.

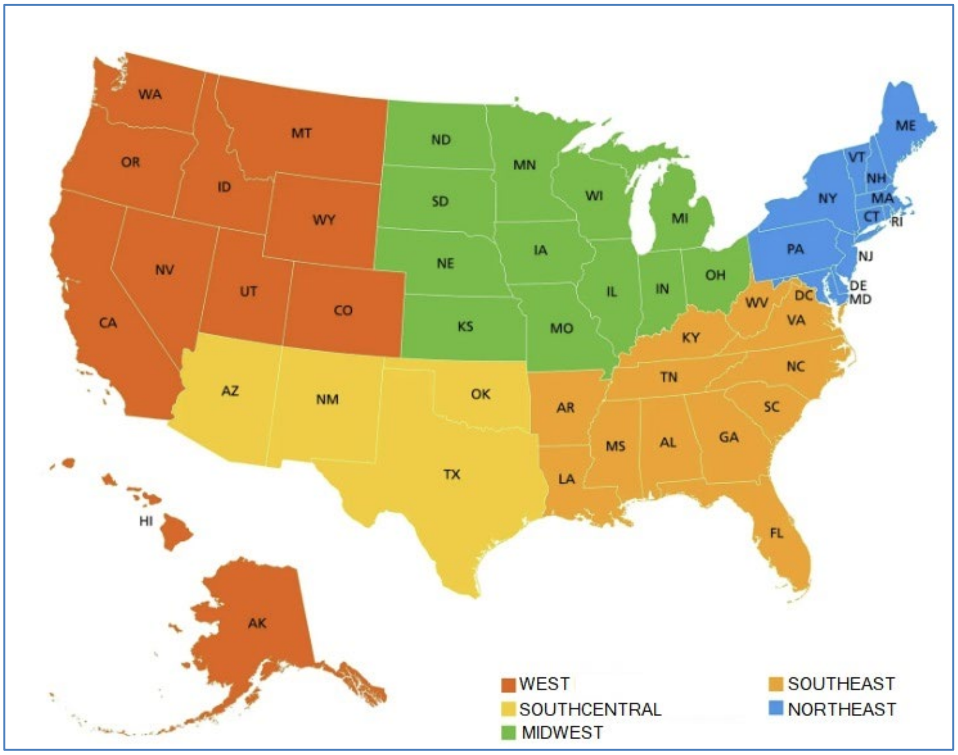


Figure 4-1 US Regions Classification Considered in this Research. (mappr.co)

4.2 Overview on the Sanitary Sewer Dataset

The sewer inspection pipe data was collected from 11 different city municipalities. This dataset from different utilities was grouped under the 4 regions namely Southeast, Southcentral, Midwest and Northeast as shown in Figure 4-1 above.

The descriptive analysis of the data variables is shown in Tables 4-1 through 4-4. The collected pipe segments have labels with names such as Pipe ID, Pipe Installation Date, Pipe Inspection Date, Pipe Material, Pipe diameter, Pipe Length, Upstream MH and Downstream MH and PACP Score. A sample-labeled dataset of this study is tabulated in Table 4-5 below. A detailed review of each of these listed variables is discussed below.

Table 4-1 Descriptive Statistics of Numerical Variables – Southeast

Independent variable	Count	Minimum	Maximum	Mean	Standard deviation
Age (years)	3054	5	103	49.49	18.62
Diameter (in.)	3054	5	90	10.5	7.62
Length (ft.)	3054	2	2261	257.58	152.99
Slope (ft.)	3054	-0.97	6.20	0.44	0.45

Table 4-2 Descriptive Statistics of Numerical Variables – Southcentral

Independent variable	Count	Minimum	Maximum	Mean	Standard deviation
Age (years)	4268	1	121	40.65	22.39
Diameter (in.)	4268	4	96	11.26	9.97
Length (ft.)	4268	0.5	2054	269.71	223.24
Slope (ft.)	4268	-0.0002	14.12	0.66	1.17

Table 4-3 Descriptive Statistics of Numerical Variables –Midwest

Independent variable	Count	Minimum	Maximum	Mean	Standard deviation
Age (years)	4268	1	121	40.65	22.39
Diameter (in.)	4268	4	96	11.26	9.97
Length (ft.)	4268	0.5	2054	269.71	223.24
Slope (ft.)	4268	0.002	14.12	0.66	1.17

Table 4-4 Descriptive Statistics of Numerical Variables –Northeast

Independent variable	Count	Minimum	Maximum	Mean	Standard deviation
Age (years)	2088	1	107	53.75	19.07
Diameter (in.)	2088	6	96	22.33	14.36
Length (ft.)	2088	1	1238	219.08	148.75
Slope (ft.)	2088	0.000	0.59	0.12	0.17

Table 4-5 Sample Dataset Collected for the Research

Pipe ID	Pipe Inspection Date	Pipe Construction Date	Pipe Diameter	Pipe Material	Pipe Length	Upstream MH	Downstream MH	PACP Score
1	4/28/2022	7/25/1958	8	DIP	398.9	400761	400350	2
2	8/22/2022	8/18/1988	8	PVC	239.1	400880	400837	3
3	12/17/2021	7/17/2001	6	VCP	137.8	402985	402984	1
4	12/16/2021	7/1/2004	10	PVC	170.7	400872	400875	1
5	8/17/2022	12/3/1964	12	DI	255.6	401277	403390	3
6	12/13/2021	12/9/1964	8	PVC	242.8	401300	401301	3
8	8/15/2022	3/1/2005	10	PVC	177.9	400792	400791	1
9	1/4/2022	4/25/2002	10	UNREINCONC	63.8	401334	401337	1
10	8/23/2022	2/28/2002	8	PVC	175.2	400978	400977	4
11	1/10/2022	6/11/1947	8	PVC	208.8	400586	400585	1
12	5/3/2022	5/10/1965	10	AC	137.7	401871	401872	3
13	8/15/2022	6/5/1975	8	PVC	312.5	402438	402042	5
14	12/13/2021	8/6/1985	12	PVC	191.3	401034	401035	1
15	2/11/2022	9/7/2001	8	RCP	60.2	403295	403296	2
16	5/2/2022	6/8/1999	15	AC	16.3	408036	401887	2
17	8/24/2022	7/3/2001	10	CP	26.2	401279	403390	3
18	4/18/2022	8/5/2006	12	AC	174.9	401347	401346	1
19	2/10/2022	4/2/2005	6	PVC	126	400545	400560	5
20	5/10/2022	2/3/1998	8	RCP	361	401340	401315	3

4.2.1 Preparation of Data

Data preparation is a collection of techniques for manipulating datasets to submit raw information as input into modeling algorithms, thereby obtaining greater precision. Data preparation is required before constructing statistical or artificial intelligence models. Several techniques should be used to prepare the dataset, as data preparation is not an entirely automated procedure (Pyle, 2007). Before beginning statistical analysis of the sewer dataset, data were filtered, and several evaluations were conducted to identify incorrect and absent information. Each pipe segment was assigned a unique "Utility ID" to simplify identifying and tracking individual pipes. In addition, these utility IDs were used to identify any duplicate pipe dataset records.

The sewer dataset included several missing values on installation year information, pipe material, pipe slope, length, and condition scales. About 2000 pipe segments were removed manually from the dataset on the excel sheet because of the missing values. Negative slope and length values were extracted. Low-population pipe materials, including plastic pipe, Polypropylene, and segmented blocks, were removed, totaling approximately 500 pipes.

Boxplots were used to remove outliers from the dataset as a final phase. Observed datasets frequently include outliers that are numerically distinct from the remaining data. Typically, outliers are larger or smaller than the experimental values in the dataset. Boxplot is a well-known, straightforward graphical representation of continuous data variation. Boxplot identifies the median, the lower quartile, the upper quartile, the lower extreme, and the upper extreme (Seo, 2006). After treating the outliers, it was observed that the correlation between dependent and independent variables improved. The final dataset consists of distinct pipe segments with various physical and environmental variables, as shown in Table 4-6. Information on the pipe variable division is illustrated in Table 4-7. This table is adapted from the Malek Mohammadi (2019).

Table 4-6 Variables Included in Sewer Pipe Dataset (Adopted from Malek Mohammadi, 2019)

Category	Variables	Description
Physical	Age	Time difference between the installation date of the pipe segment and date of inspection in years
	Material	Type of sewer pipes material
	Diameter	Diameter of the sewer pipe segment in inches
	Slope	Vertical displacement of the pipe section per horizontal displacement in percentage
	Length	Length of the sewer pipe segment between two manholes in feet
Environmental	Soil Type	Type of soil surrounding the pipe
	US Region	Area division according to the US region

Table 4-7 Information on Variable type in the Dataset

Variables	Variable Type
Age	Continuous quantitative
Diameter	
Slope	
Length	
Material	Nominal categorical • VCP • RCP

Soil Type	Nominal categorical • Clay • Sand • Silt

4.2.2 Pipe Age

The pipe age for this study was calculated by using the pipe construction year and inspection year from the data. It was remarkable to learn that sewer pipelines were installed as early as 1901 and are still used today. Most pipelines installed at the beginning of the twentieth century were made of concrete or vitrified clay. Figure 4-2 presents the overview of the pipe age for all Southeast, Southcentral, Midwest, and Northeast. It can be observed that 75% of pipes are in the age range between 20 years to 70 years. The oldest pipe installed for Southeast, Southcentral, Midwest, and Northeast are in the year 1919, 1901, 1915 and 1903 respectively and the sewer inspection year considered for the regions is 2022.

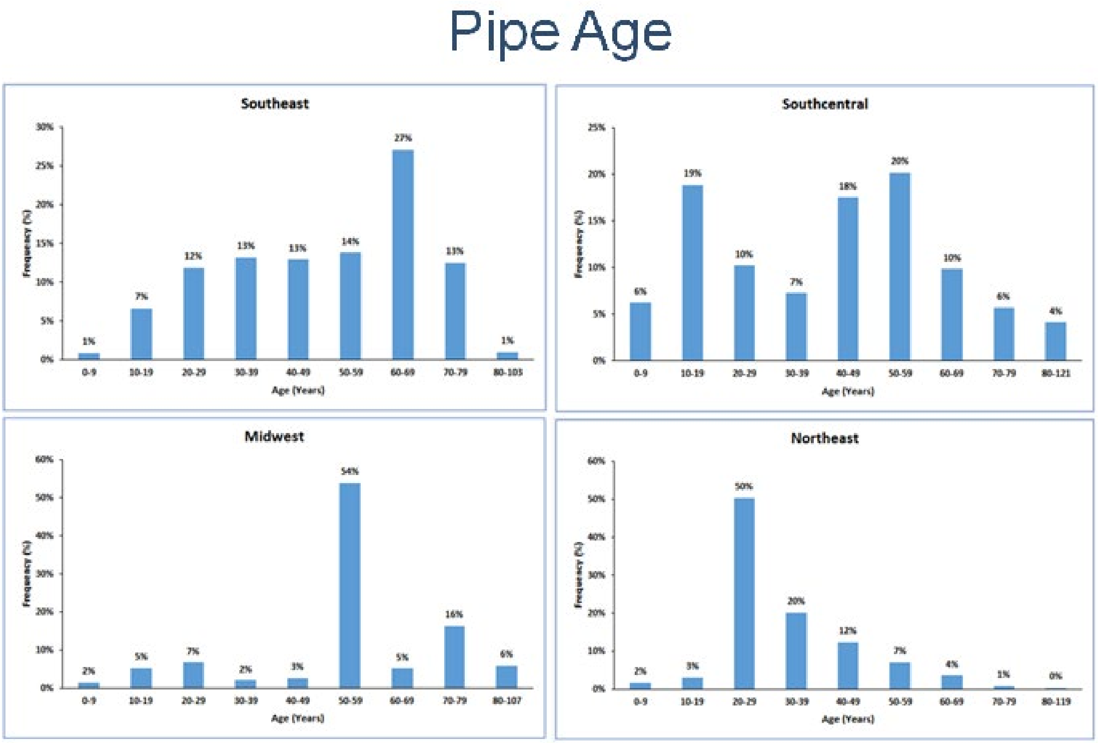


Figure 4-2 Overview of Pipe Age Frequency for all the Regions Under Study

4.2.3 Pipe Material

Figure 4-3 illustrates the pipe materials details under each of US region considered under this research. The names of the pipes included are Vitrified clay pipe (VCP), Polyvinyl Chloride (PVC), Reinforced Concrete Pipe (RCP), Unreinforced Concrete Pipe (UnReinCONC), Reinforced Polymer Mortar (RPM), OTHERS (pipes with unknown material), Ductile Iron (DI), Cast Iron (CI), Asbestos Cement (AC), Fiberglass Reinforced Pipe, High-Density Polyethylene (HDPE), Pre-stressed Concrete Cylinder Pipe (PCCP), Clay-lined Concrete Pipe (CLC), Corrugated Metal Pipe (CMP). It can be seen in Figure 4-3 below that, in all regions many pipes are made of VCP and PVC. And on the other hand, AC, DI, CI and HDPE are used rarely.

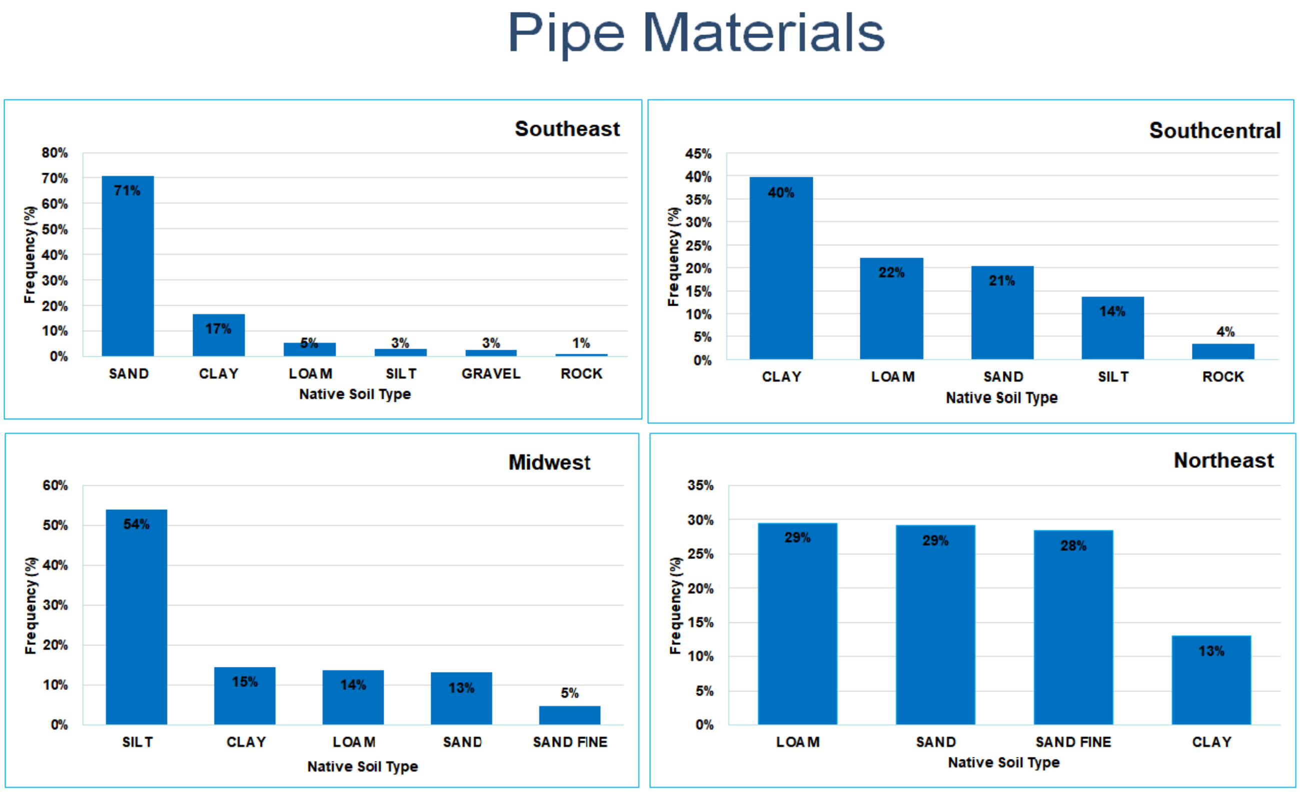


Figure 4-3 Overview of Pipe Material Frequency for all the US Regions

4.2.4 Pipe Diameter

Figure 4-4 illustrates the pipe diameter details under each US region considered in this research. The range of diameter in the entire dataset was found to be 4 to 96 inches. It can be observed that most of the pipes are between 4 to 10 inches.

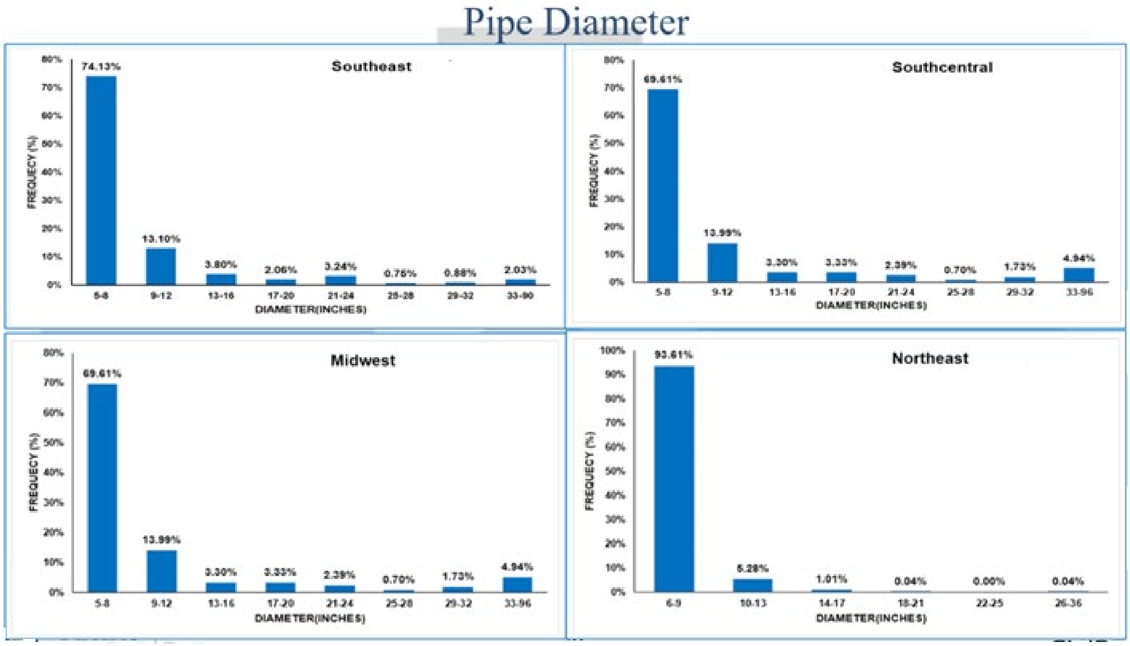


Figure 4-4 Overview of Pipe Diameter Frequency for all the US Regions

4.2.5 Pipe Soil Native Type

The soil type is one of the most significant factors affecting ground stability and sewer pipe stability. The soil type collected for this study is the soil which is soil used to backfill the trench after the pipe installation, there is no information whether this soil type is used for providing the support to the pipe. Hence in this research this soil is assumed to be the native soil surrounding the pipes underground. The soil type is one of the independent variables considered in this model building which denotes the type of material used to backfill around sewer pipes. The soil type found to be included in the dataset includes sand, silt, loam, clay, sand fines, rock, gravel. Figures 4-5 show the overview of pipe diameter frequency for all the US Regions.

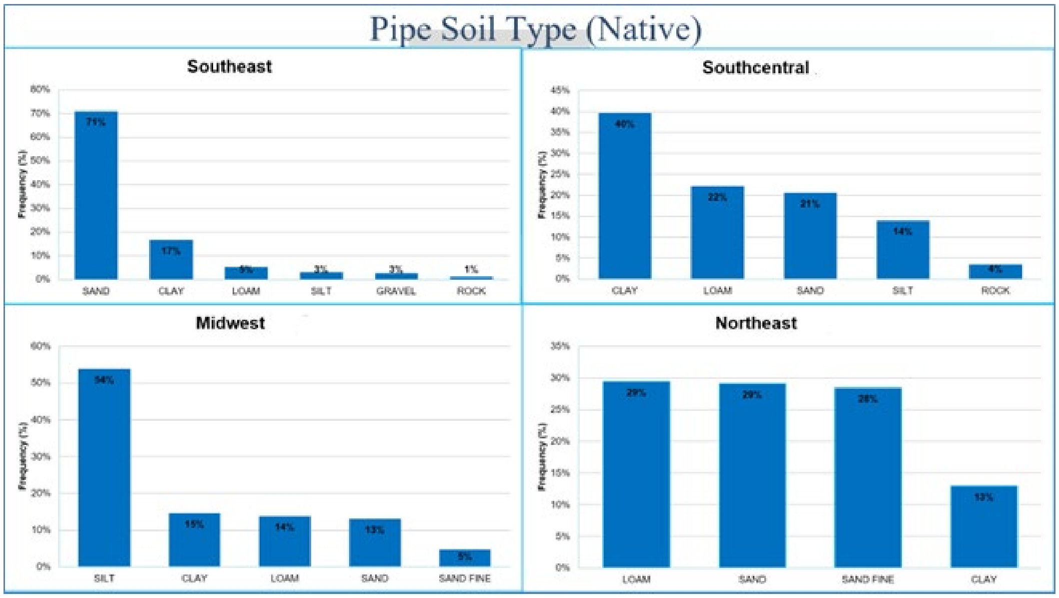


Figure 4-5 Overview of Pipe Native Soil Type Frequency for all the US Regions

4.2.6 Pipe Length

Pipe length is measured in feet from manhole to manhole for sewer pipe segments. Figure 4-6 show the overview of the pipe length details included in this research for all the US regions. It can be observed in the below figure, the longest pipe in the dataset considered under this study is about 2240 feet and the smallest pipe length is found to be 0.5 feet.

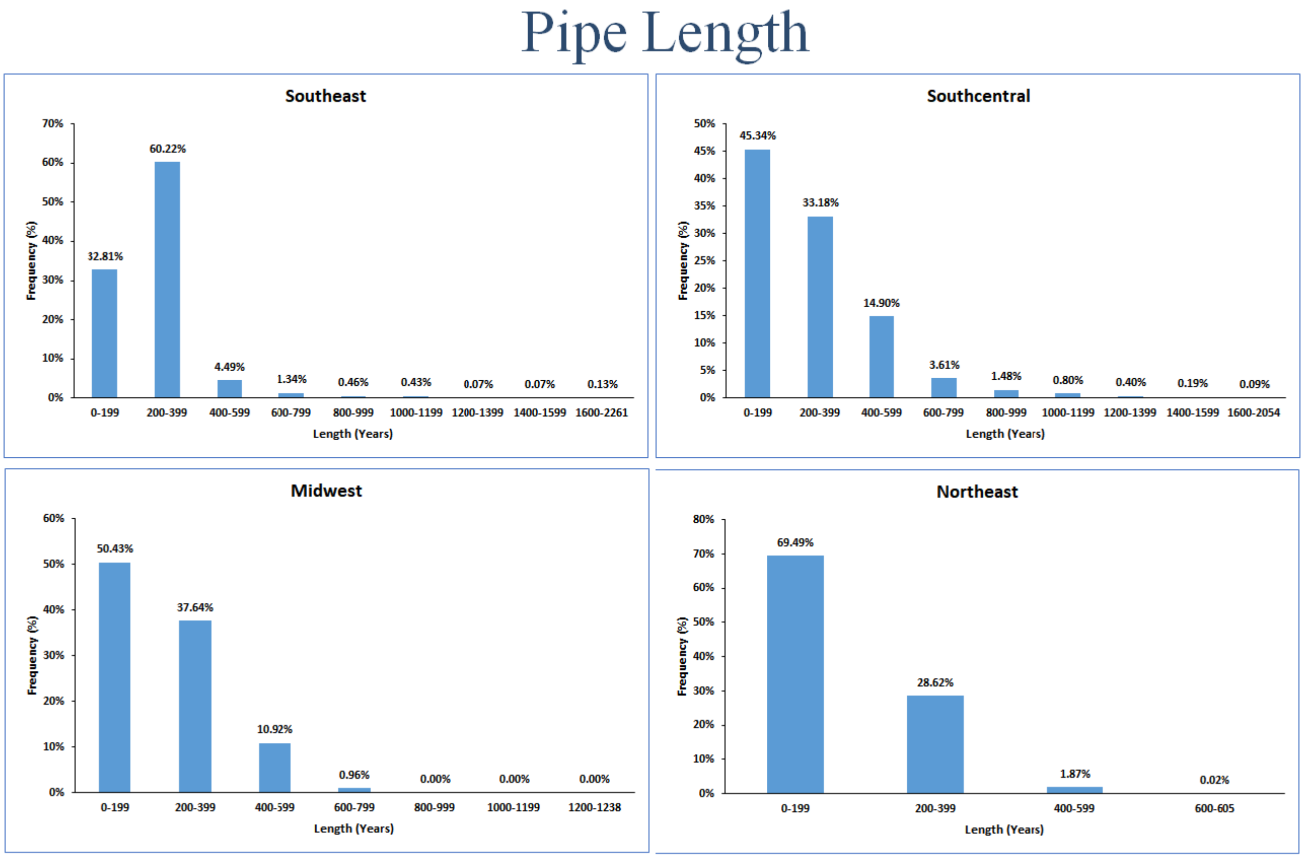


Figure 4-6 Overview of Pipe Length Frequency for all the US Regions

4.2.7 Pipe Slope

The slope of a pipe is the ratio of the vertical displacement of a pipe section to its horizontal displacement measured in feet. The pipe slope was calculated using the difference between the upstream and downstream elevation values. Figure 4-7 show the overview of the pipe slope frequency for the pipes considered under each US region. And it can be observed that most of the pipes have slopes between 0.5-1.0 feet.

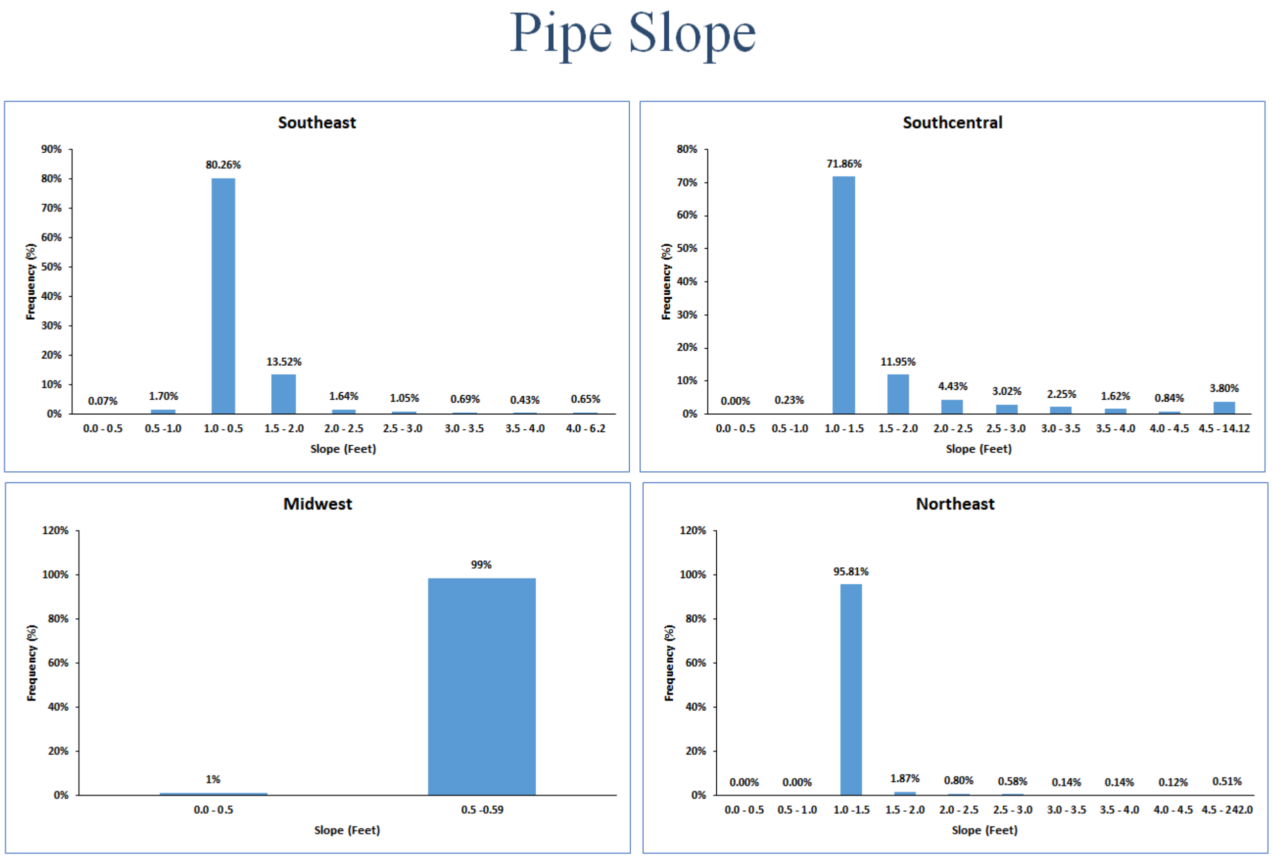


Figure 4-7 Overview of Pipe Slope Frequency for all the US Regions

4.3 Correlation Analysis

Correlation analysis is an approach to statistics for measuring the degree of association between two variables. There are occasions when there is no correlation between two variables. A significant relationship indicates that one variable's value can be anticipated based on the other variable's value. Conversely, when the relationships between variables are feeble, they cannot be accurately predicted. The coefficient of correlation within variables can be either positive or negative but must fall within the range of -1.00 and 1.00. Combining significantly correlated independent variables into the model may cause a multicollinearity issue that impacts the model's results. It is not recommended to develop a model with highly correlated independent variables. In the model, most variables in the data set accessible were not normally distributed. Thus, spearman's rank correlation was employed to investigate the connection between the variables. Spearman's rank correlation can be used to characterize the relationship between variables that are not linearly related. This method makes no presumptions about the distribution of the model's variables, unlike Pearson's method, which implies the normal distribution of two variables and can only describe the linear relationship between two variables. Figure 4-8 to Figure 4-11 illustrates correlation of attributes for all the US regions.

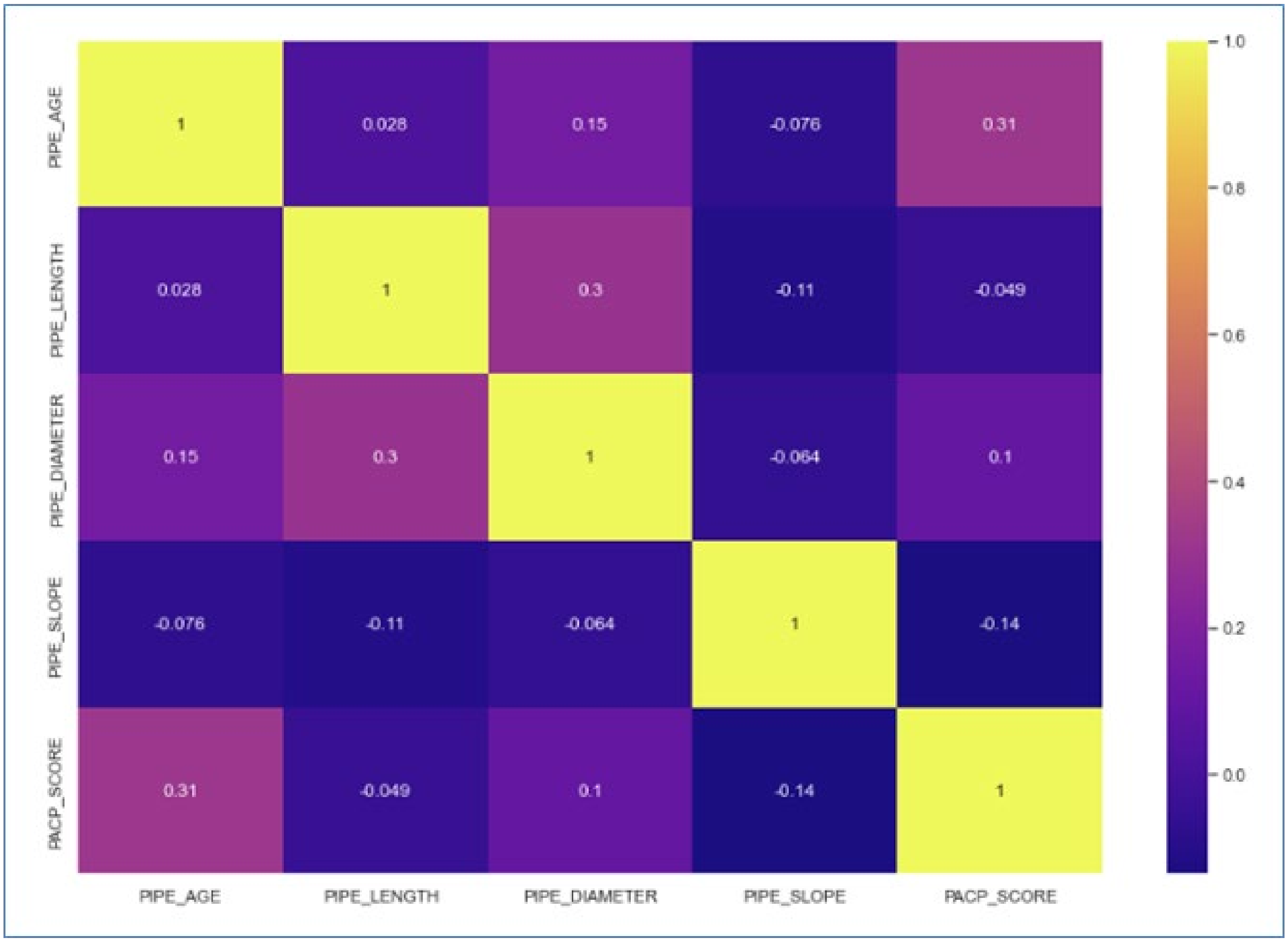


Figure 4-8 Correlation of Attributes-Southeast

In the above Figure 4-8, there is highest correlation between pipe age and PACP score (+0.31). Also, there is no strong correlation among the variables which are independent which indicates that none of them needs to be eliminated from the model to avoid multicollinearity.

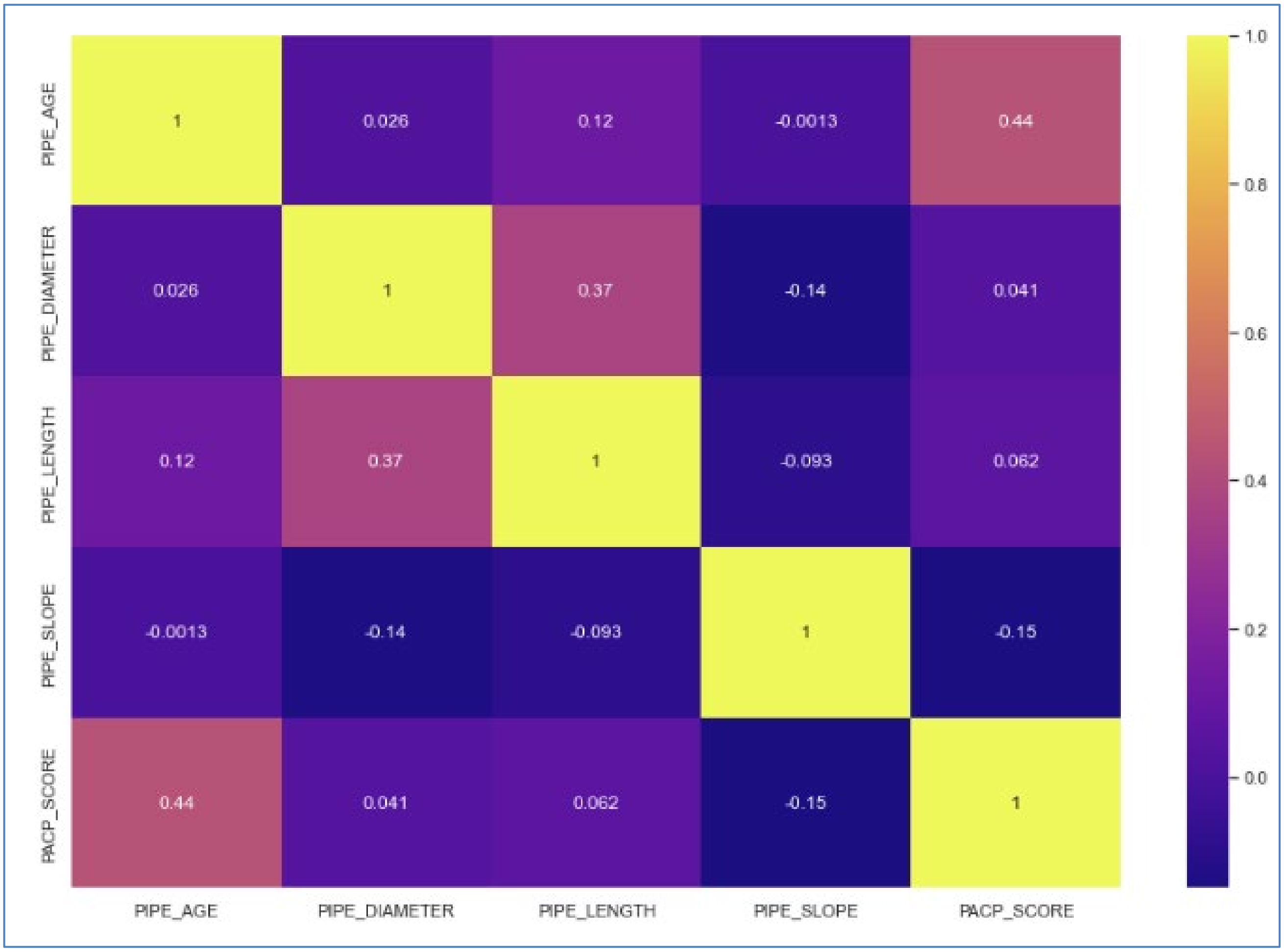


Figure 4-9 Correlation of Attributes-Southcentral

In the above Figure 4-9, there is highest correlation between pipe age and PACP score (+0.44). Also, there is no strong correlation among the variables which are independent which indicates that none of them needs to be eliminated from the model to avoid multicollinearity.

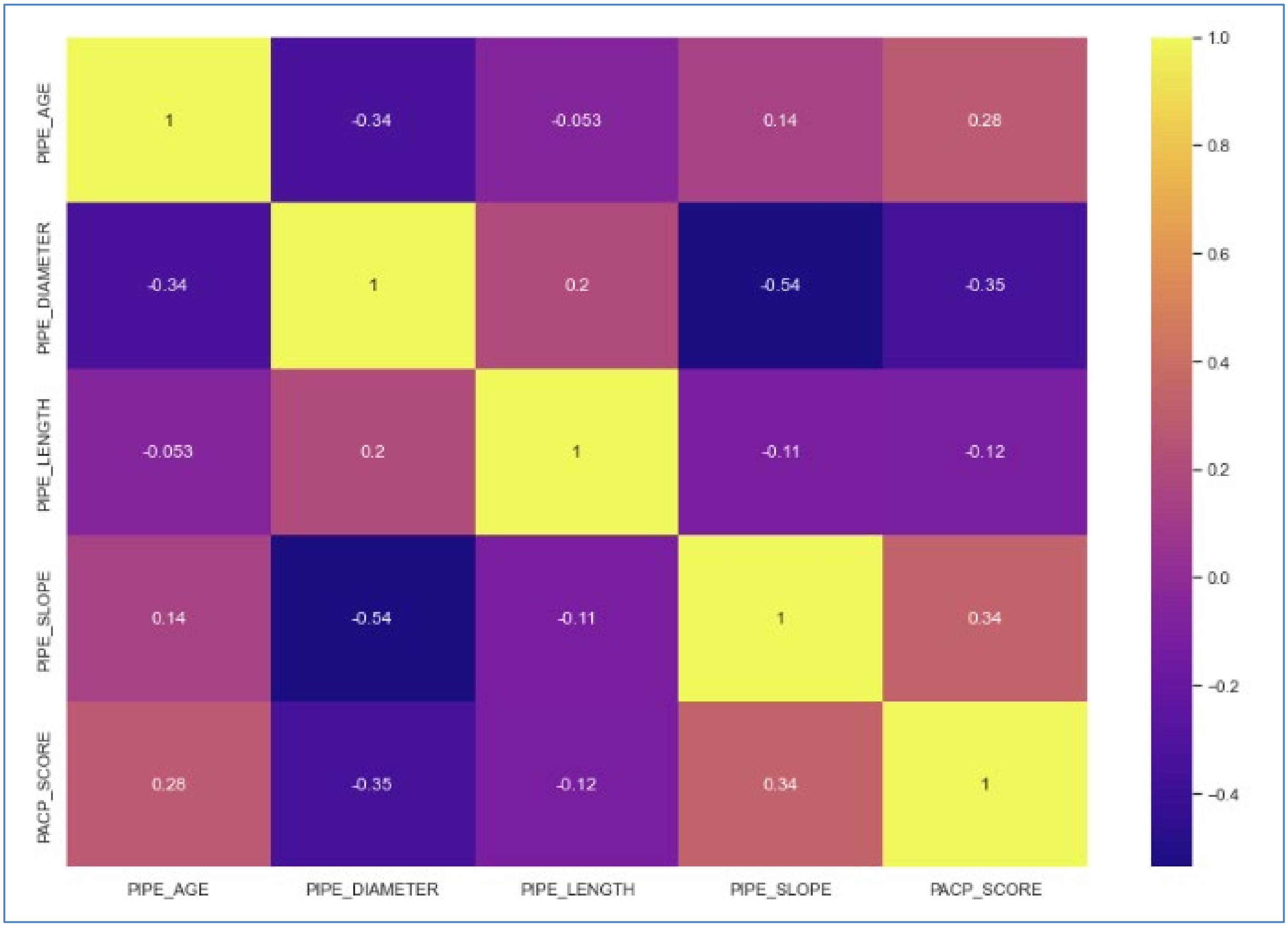


Figure 4-10 Correlation of Attributes-Midwest

In the above Figure 4-10, there is highest correlation between pipe slope and PACP score (+0.34) and pipe age and PACP score has good correlation (+0.28), Also, there is no strong correlation among the variables which are independent which indicates that, none of them needs to be eliminated from the model to avoid multicollinearity.

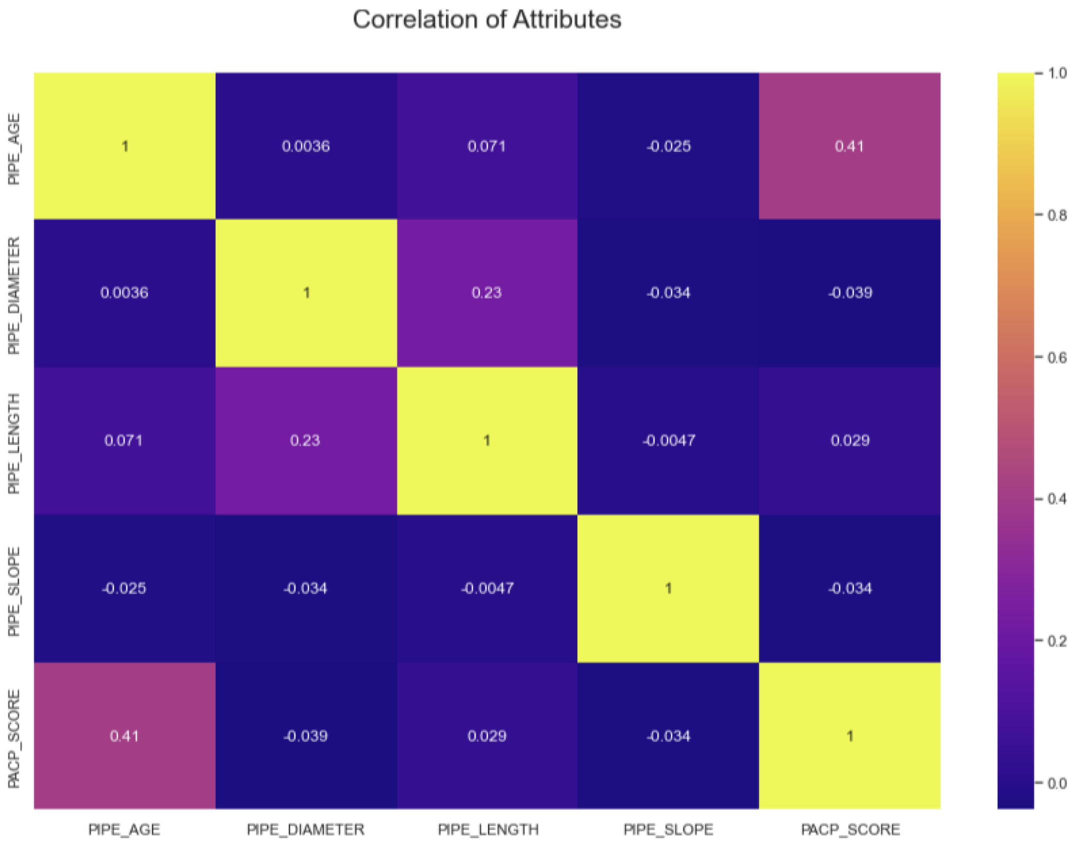


Figure 4-11 Correlation of Attributes – Northeast

In the above Figure 4-11, there is highest correlation between pipe age and PACP score (+0.41), Also, there is no strong correlation among the variables which are independent which indicates that, none of them needs to be eliminated from the model to avoid multicollinearity.

4.4 Chapter Summary

In this chapter, the source of the dataset for sanitary sewers was examined in depth. In addition, the details of model variables were discussed. The original data collected was converted into a standard format in preparation for model development. The available variables for the development of the model were identified, and their significance was examined graphically. Next chapter will cover a detailed explanation on all the machine learning algorithms used in development of the prediction model.

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1 Introduction

In this part of the report, prediction models are developed by grouping the dataset with respect to their regions – Southeast, Southcentral, Midwest, and Northeast. This grouping is necessary as the sewer pipe data is collected from different utilities around the US regions, which exhibit variations in environmental conditions, soil type, construction and installation methods, levels of inspection, pipe material types, and other conditions. Therefore, models are developed based on each US region mentioned above, and their performances are compared to see the practical variations of the independent variable on the target variable, the PACP score.

The models are designed using logistic regression tree-based models, which includes a decision tree classifier, random forest classifier, Adaboost classifier, gradient boost classifier, and XGBoost classifier based on their concepts as discussed in Chapter 4.

The exploratory data analysis of this study reveals that PACP scores of 4, and 5 have significantly fewer instances than the scores 1, 2 and 3, this shows that the dataset is imbalanced and leads to a poorly trained model performance. When classification methods employ imbalanced data, trained models may perform inadequately. This imbalance data condition needs to be handled more cautiously as misclassifying the data with minority scores with that of majority score class would cause worse impact on the model performance. In this study, over-sampling technique is utilized to treat the imbalanced data. Models are then developed using imbalanced and over-sampled data and their performance is presented under each model results section. Table 5-1 presents the basics of the evaluation metrics which are used in classification of the machine learning algorithms. Chapter 4 has a detailed explanation of the evaluation metric.

Table 5-1 Basic Evaluation Metric Considered in this Study.

Evaluation metric	Significant	Not- Significant
Confusion Matrix	✓	
ROC Curve	✓	
AUC	✓	
Accuracy	✓	
Recall	✓	
Precision		×
F1- Score	✓	

5.2 Performance of Developed Models Based on US Regions

5.2.1 Southeast

5.2.1.1 Dataset Summary

Table 5-2 presents the descriptive statistics of variables such as pipe age, pipe diameter, pipe length, and pipe slope for the dataset collected from the Southeast.

Table 5-2 Descriptive Statistics of Variables in Southeast.

	Pipe Age	Pipe Diameter	Pipe Length	Pipe Slope	PACP Score
count	3054	3054	3054	3054	3054
mean	49.49	10.57	257.59	0.44	2.54
std	18.62	7.62	152.98	0.45	1.08
min	5	5	2	-0.97	1
25%	35	8	168.32	0.29	1
50%	56	8	258.06	0.39	3
75%	65	10	321.63	0.45	3
max	103	90	2261	6.2	5

Pipe material considered included – VCP, PVC, UnReinCONC, RCP, AC, DI, RPM, CI, CLC. The full form of the pipe material is presented in Appendix Section A. The soil type included in the dataset includes sand, silt, clay, loam, gravel, rock. Table 5-3 shows the unique values of the independent variable of the Southeast dataset. The dataset for developing the models was divided as 70% for train set and 30% for test set. The imbalanced data is treated using the oversampling SMOTE analysis.

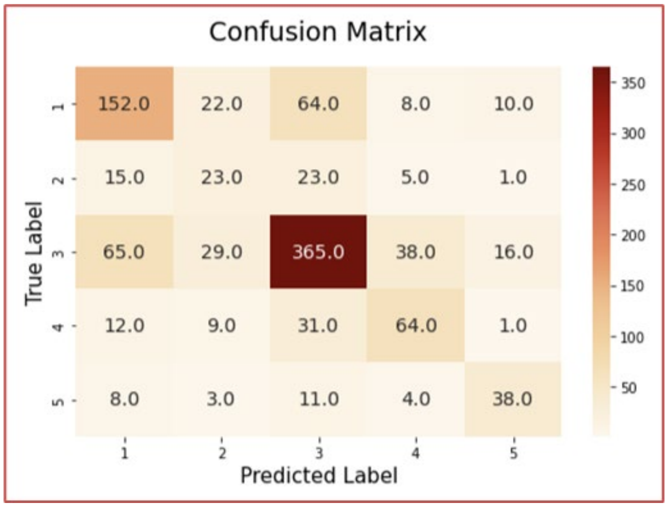
Table 5-3 Unique values of the Southeast Dataset

Pipe Age	88
Pipe Diameter	22
Pipe Material	10
Pipe Length	542
Pipe Slope	515
Pipe Slope	6
PACP Score	5

5.2.1.2 Results on Developed Models

5.2.1.2.1 Decision Tree (DT) Classifier with Imbalanced Dataset

A confusion matrix is the effective evaluation metric in model building for classification methods. The confusion matrix developed for the DT classifier method for the test set of this study is shown in Figure 5-1. The last two columns have lower data points predicted in the class of 4 and 5 scores. This is by the fact that the dataset in that class 4 and 5 are low in number when compared to the class 1, 2, 3. The ROC curve for the DT classifier is plotted and shown in Figure 5-2 and it shows that the classes 1 and 3 have good ROC score above 0.69 and the ROC for classes 2, 4 and 5 has poor score below 0.55 this is of the fact of high misclassifications in the those classes. The training set and test set performance is viewed in Table 5-4, and it is seen that the model displays poor accuracy and F1-score.



Confusion Matrix

True Label \ Predicted Label	1	2	3	4	5
1	152.0	22.0	64.0	8.0	10.0
2	15.0	23.0	23.0	5.0	1.0
3	65.0	29.0	365.0	38.0	16.0
4	12.0	9.0	31.0	64.0	1.0
5	8.0	3.0	11.0	4.0	38.0

Figure 5-1 Confusion Matrix for Decision Tree Classifier–Southeast

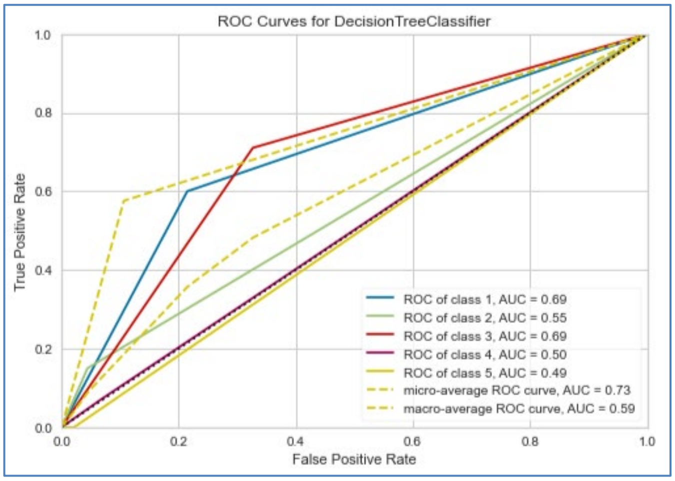


Figure 5-2 ROC Curve for Decision Tree Classifier-Southeast

Table 5-4 Performance for Decision Tree Classifier-Southeast

Performance	Accuracy	Macro F1-score
Training set	99.12%	99.23%
Test set	57.85%	55.88%

According to the ROC curve in Figure 5-2, the ROC for the classes 2, 4,5 have for It can be seen from the above Table 5-4 that the model is overfitting and has poor test accuracy of 57.85% and macro F1-Score of 55.88%.

5.2.1.2.2 Random Forest (RF) Classifier with Imbalanced Data

Random forest classifier method is under the division of bagging algorithm used as one of the best in predicting classification problems. It constructs distinct decision trees and arrives at the best result by taking the average of the results of the trees. The confusion matrix for test set of this method is shown in Figure 5-3 below and can be seen that, like DT classifier, the last two columns have less predicted values in the class of 4 and 5. The ROC curve for the RF Classifier is plotted and shown in Figure 5-4, and classes 1 and 3 have better ROC value than 2,4 and, 5 classes. The training set and test set performance is viewed in Table 5-5 and the accuracy and macro F1- score is better than DT classifier. But the model is not good as it is overfitting, and the performance is less than 70%.

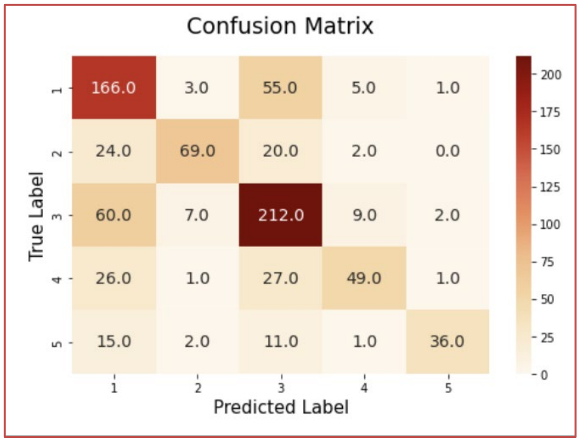


Figure 5-3 Confusion Matrix for Random Forest Classifier–Southeast

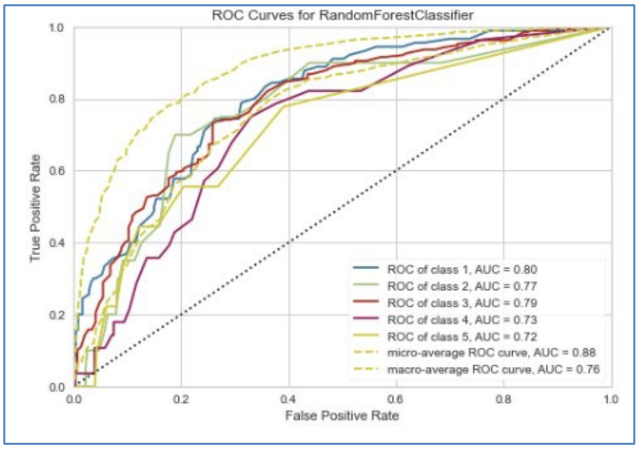


Figure 5-4 ROC Curve for Random Forest Classifier-Southeast

Table 5-5 Performance of Random Forest Classifier-Southeast

Performance	Accuracy	Macro F1-score
Training set	98.00%	98.05%
Test set	68.85%	67.23%

5.2.1.2.3 Tuned Decision Tree (DT) Classifier with Over-sampled Data

In this method of tuning the decision tree, the following grid parameters are considered and are listed in Table 5-6. Also, an approach of oversampling the dataset is carried out to treat the imbalanced data in the class of 4 and 5 PACP scores by adopting SMOTE technique, as the performance of DT model and RF model in the section 5.2.1.2.1 and 5.2.1.2.2 are not satisfactory, both over-sampling and tuning help in optimizing the given dataset. The confusion matrix of test set for this model is presented in Figure 5-5, and it can be observed that the values on diagonal element has increased and the misclassification in the classes of 4and 5 has decreased in number compared to the confusion matrix developed in the section 5.2.1.2.1. In addition to this, The ROC curve for the tuning DT Classifier is plotted and shown in Figure 5-6, and all the classes have good ROC values above 0.70. The training set and test set performance is viewed in Table 5-7 and the model has performed better than the one under section 5.2.1.2.1 and is overfitting.

Table 5-6 Tuned Decision for Tree Classifier Grid Parameters

max_depth	35
max_leaf_nodes	15
min_samples_leaf	11
random_state	1

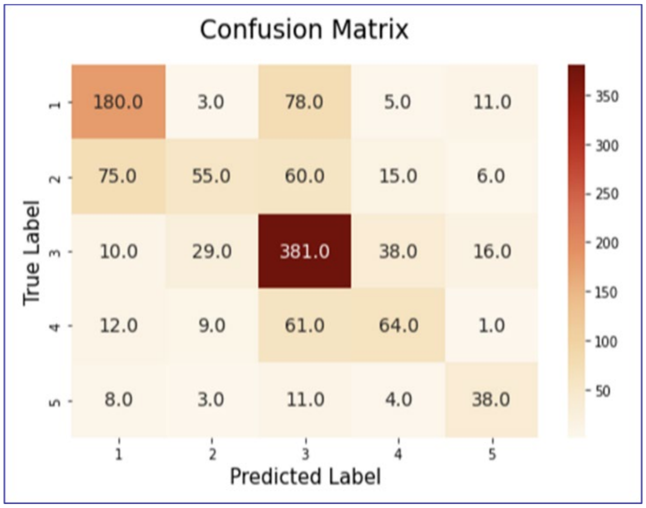


Figure 5-5 Confusion Matrix for Tuned Decision Tree Classifier–Southeast

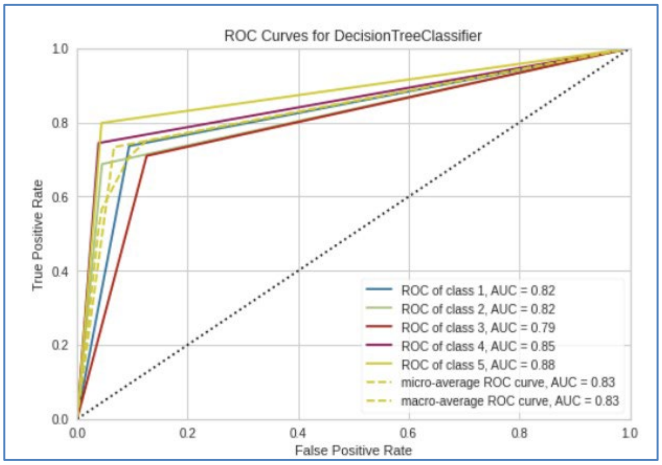


Figure 5-6 ROC Curve for Tuned Decision Tree Classifier-Southeast

Table 5-7 Performance of Tuned Decision Tree Classifier-Southeast

Performance	Accuracy	Macro F1-score
Training set	74.00%	75.12%
Test set	63.02%	60.03%

The performance of the tuned DT classifier shows that model is overfitting and has a poor F1-score of 60.03%.

5.2.1.2.4 Tuned Random Forest (RF) Classifier using Oversampled Data

Tuning of RF classifier is developed with grid parameters is shown in Table 5-8 to improve the RF classifier performance build with default parameters as its performance was unsatisfactory. During the training of the RF model, the best hyperparameters were set by trial-and-error method to obtain the best possible performance. The confusion matrix of test set for the tuned RF classifier is presented in Figure 5-7 below. The predicted values on diagonal are found to increase when compared to the RF model with default parameters and the misclassification has reduced in all the classes. The ROC curve for the tuning RF Classifier is plotted and shown in Figure 5-8 and classes 1 and 5 have good ROC values and class 2 has the least ROC value. The training set and test set performance for this model on accuracy and macro F1- Score is viewed in Table 5-9. This model is not overfitting and has better performance with an F1-score of 73.32%.

Table 5-8 Tuned Random Forest Classifier Grid Parameters

max_depth	10
max_samples	0.2
min_impurities_decrease	0.0001
N_estimates	150
random_state	1

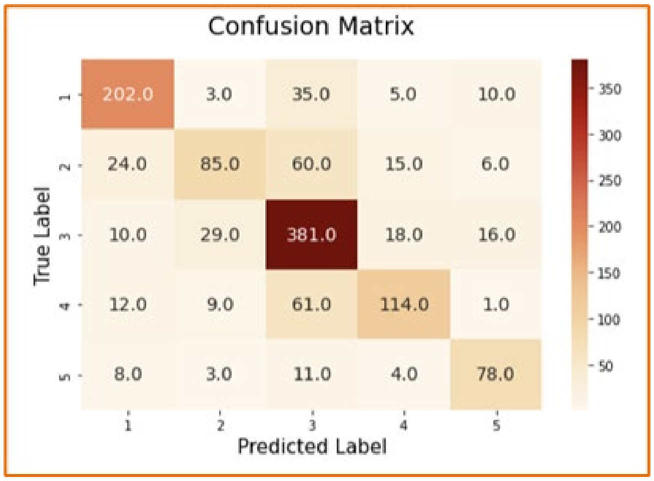


Figure 5-7 Confusion Matrix for Decision Tree Classifier–Southeast

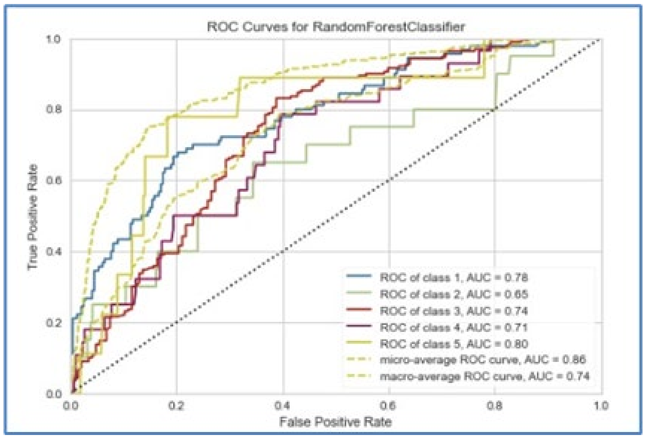


Figure 5-8 ROC Curve for Tuned Random Forest Classifier-Southeast

Table 5-9 Performance of Tuned Random Forest Classifier-Southeast

Performance	Accuracy	Macro F1-score
Training set	76.02%	75.11%
Test set	75.42%	73.32%

5.2.1.2.5 AdaBoost Classifier with Default Parameters using Imbalanced Data

AdaBoost Classifier with default parameter algorithm is one of the ensembles boosting algorithms employed in model development of this research. AdaBoost determines predictions from every predictor and weights them according to their predictor weights (the higher the predictor weight, the more accurate the predictor). The predicted class is determined by many weighted ballots (Shirkhanloo, 2022). The two most important parameters of boosted models are the number of trees and the learning rate, which determines how much each tree is permitted to rectify the errors of the previous trees. These two parameters are related because constructing a model with the identical level of complex at a slower learning rate requires more trees. In this model development, the number of predictors were set to 50 and learning was taken as 1. The confusion matrix for this model is shown in Figure 5-9. The actual predicted values generated in all the classes are found to have better values than RF model with default parameters. The ROC curve for the AdaBoost classifier is plotted as in Figure 5-10. The training set and test set performance for accuracy and macro F1- score is viewed in Table 5-10.

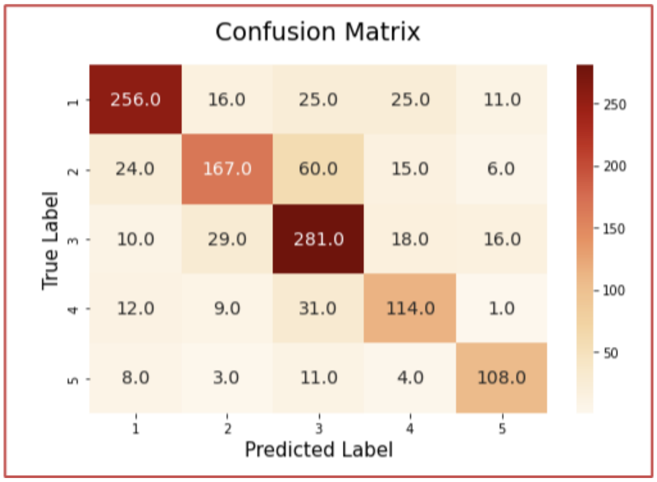


Figure 5-9 Confusion Matrix for Adaboost Classifier–Southeast

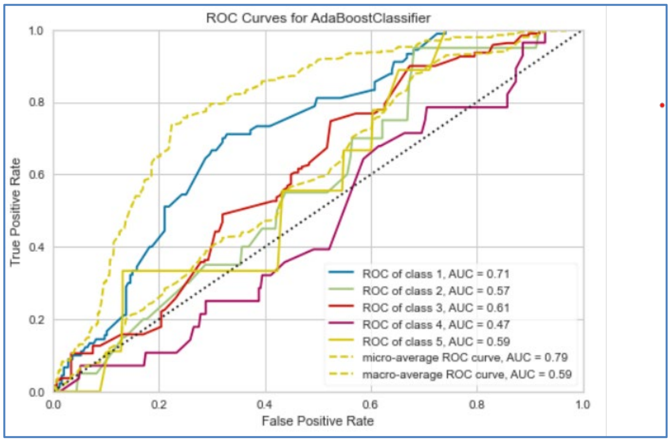


Figure 5-10 ROC Curve for AdaBoost Classifier-Southeast

The AUC for classes 2,3,4 and 5 is low as seen in Figure 5-10, this indicates that the predicted value in every class is not good. The performance of this model is shown in Table 5-10 and F1-score is calculated as 58.25% which is not satisfactory.

Table 5-10 Performance of AdaBoost Classifier-Southeast

Performance	Accuracy	Macro F1-score
Training set	69.19%	69.17%
Test set	59.13%	58.25%

5.2.1.2.6 Gradient Boost Classifier with Default Parameters using Imbalanced Data

Another boosting algorithm used in this research is gradient boosting classifier with default parameter. Gradient Boosting operates likewise to AdaBoost by in sequence adding predictors to an ensemble, with each one improving its predecessor. But unlike AdaBoost, the gradient boost 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). The confusion matrix for test set of this model is shown in Figure 5-11 and it illustrates that the values on diagonal element are predicted rightly and the misclassification in all the classes are improved compared to AdaBoost model. The ROC curve for this model is plotted Figure 5-12. The ROC curve depicts that, the AUC for all classes have good score above 0.7 indicating that model has developed in good manner and will have improved performance. The training set and test set performance for accuracy and macro F1-

score is viewed in Table 5-11. This model is having test macro F1-score of 73.19% and it is not overfitting and has better scores than AdaBoost Classifier.

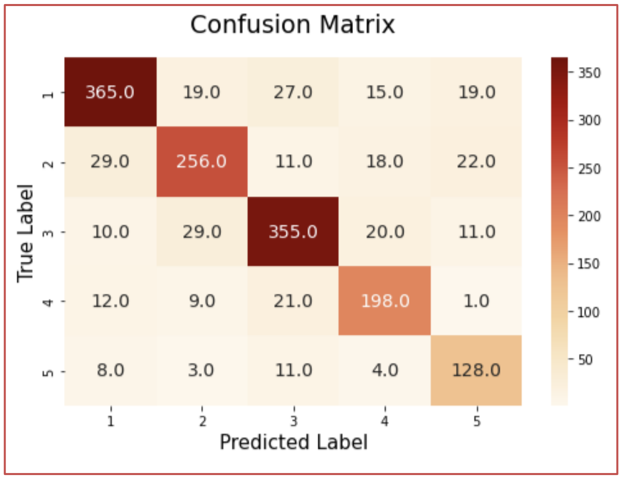


Figure 5-11 Confusion Matrix for Gradient Boost Classifier–Southeast

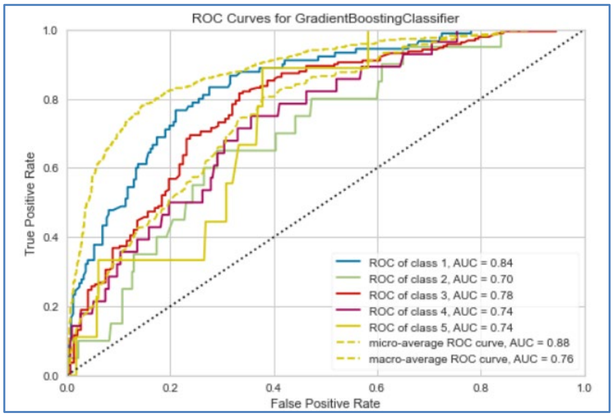


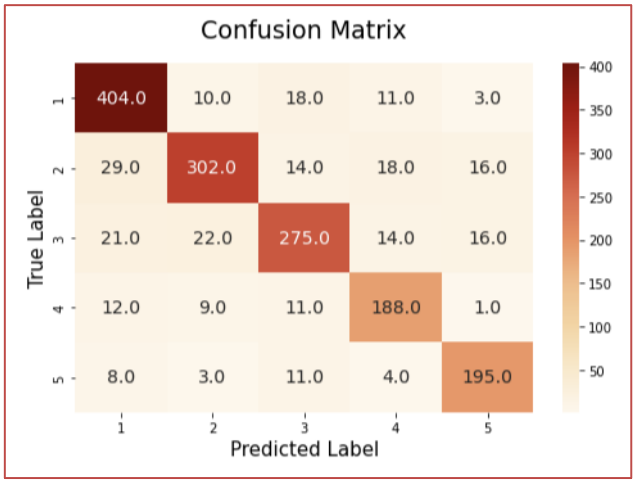
Figure 5-12 ROC Curve for Gradient Boost Classifier-Southeast

Table 5-11 Performance of Gradient Boost Classifier -Southeast

Performance	Accuracy	Macro F1-score
Training set	79.25%	78.33%
Test set	75.14%	73.19%

5.2.1.2.7 XGBoost Classifier with Default Parameters using Imbalanced Data

XGBoost is the other most effective ensemble algorithm in predicting classification. It computes better and faster than the other algorithms. XGBoost computes second partial derivatives of the loss function which is discussed in Chapter 3 in Section 3.5.3.4. The confusion matrix of test set and the ROC curve for this model is presented in Figure 5-13 and Figure 5-14. There is drastic increase in the predicted true values on diagonal element and misclassification in classes is low compared to the other tree- based models discussed earlier in this section. The training set and test set performance for accuracy and macro F1- score is listed in Table 5-12. This model has the best performance of all the other models which are built with default parameters using imbalanced data. The model is not overfitting.



True Label \ Predicted Label	1	2	3	4	5
1	404.0	10.0	18.0	11.0	3.0
2	29.0	302.0	14.0	18.0	16.0
3	21.0	22.0	275.0	14.0	16.0
4	12.0	9.0	11.0	188.0	1.0
5	8.0	3.0	11.0	4.0	195.0

Figure 5-13 Confusion Matrix for XGBoost Classifier-Southeast

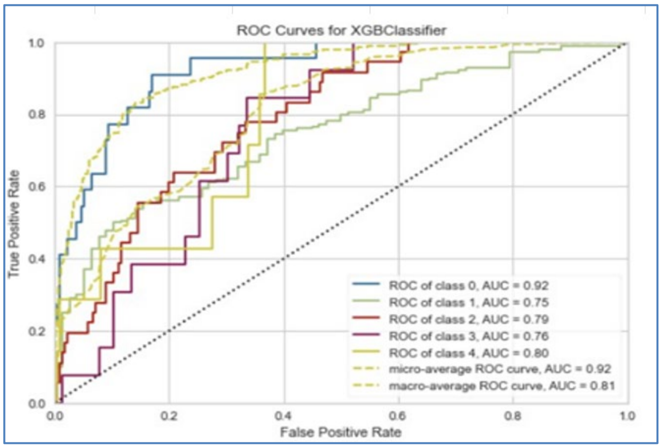


Figure 5-14 ROC Curve for XGBoost Classifier-Southeast

Table 5-12 Performance of XGBoost Classifier-Southeast

Performance	Accuracy	Macro F1-score
Training set	77.01%	79.23%
Test set	75.15%	75.20%

5.2.1.2.8 Tuned AdaBoost Classifier using Over-sampled Data

In this model, two techniques of hyperparameter tuning were carried out namely gridsearch CV and randomizedsearch CV. The Confusion matrix for both tuning methods are presented in Figure 5-15(a) and Figure 5-15(b). It can be seen that a good number of values are predicted well on the diagonal element of matrix but, also there is huge misclassification of the values in all the classes which will affect the model performance. The ROC curve for both gridsearch CV and randomizedsearch CV is shown in Figure 5-16(a) and Figure 5-16(b), it can be illustrated that AUC for all has poor score except for the class 4 which is 0.71. And the ROC curve for randomizedsearch CV presents that AUC for all 2,3 and 5 classes are low and for classes 1 and 5 is it slightly better value above 0.70. The training set and test set performance for accuracy and macro F1- Score are listed in Table 5-13 and both the models are overfitting and have poor performance below 70%.

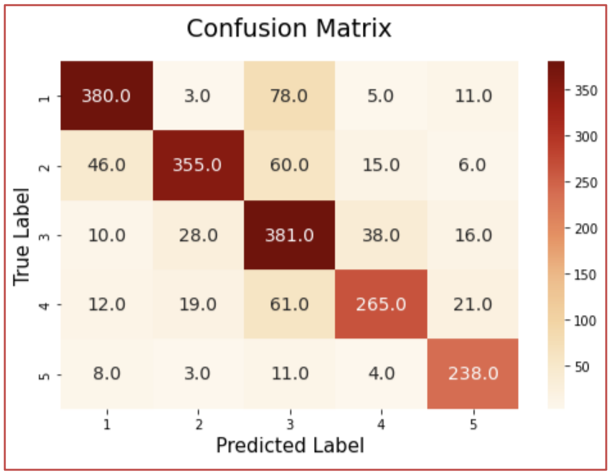


Figure 5-15 (a) Confusion Matrix for Adaboost Classifier using Gridsearch CV–Southeast

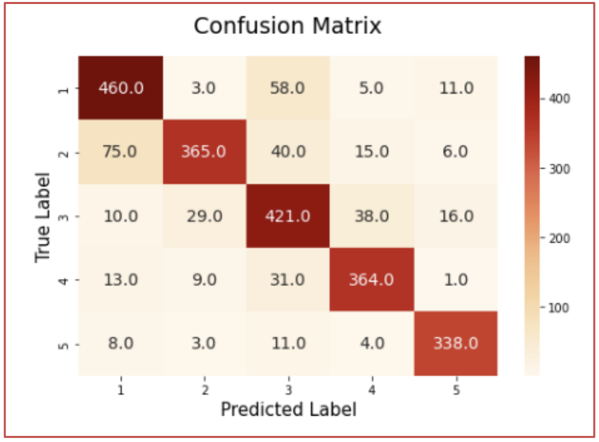


Figure 5-15 (b) Confusion Matrix for Adaboost Classifier using Randomizedsearch CV–Southeast

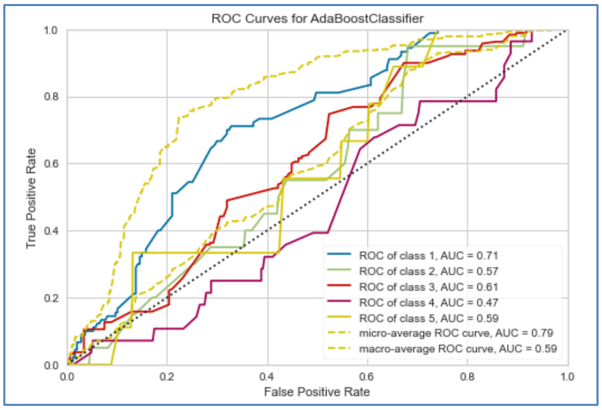
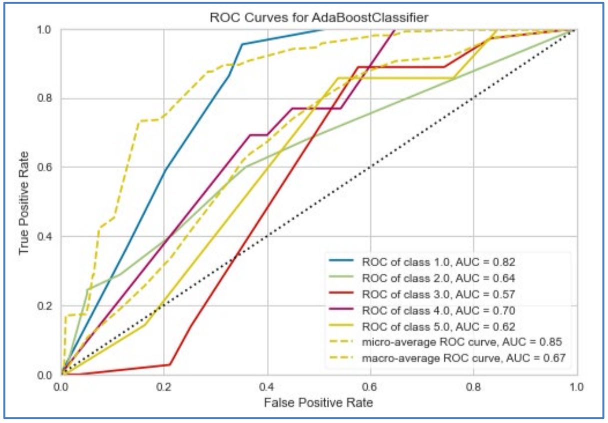


Figure 5-16 (a) ROC Curve for Tuning AdaBoost Classifier using Gridsearch CV -Southeast



5-16 (b) ROC Curve for Tuning AdaBoost Classifier using Randomizedsearch CV-Southeast

Table 5-13 Performance of Tuned AdaBoost Classifier-Southeast

Performance	Accuracy	Macro F1-score
Gridsearch CV Training set	100.00%	100.00%
Gridsearch CV Test set	64.5%	64.5%
Randomizedsearch CV Training set	100.00%	100.00%
Randomizedsearch CV Test set	63.50%	63.50%

5.2.1.2.9 Tuned Gradient Boost Classifier using Over-sampled Data

Similarly, as explained in section 5.2.1.2.8, gradient boosting algorithm was tuned with hyperparameters to optimize and evaluate the model by gridsearch CV and randomizedsearch CV. The confusion matrix is illustrated in Figure 5-17(a) and 5-17(b) below. The predicted value on the diagonal area is evidence that the model performance is rightly evaluated. For the 4 and 5 classes it can be seen that wrongly predicted values are counting to 0, which is a good, developed model. It is evident that the confusion matrix developed for these models is better than AdaBoost models as there is less misclassification in all classes and the higher values predicted rightly on diagonal element. Figures 5.18(a) and 5.18(b) represent the ROC curve for these models. Table 5-14 shows the performance of these models.

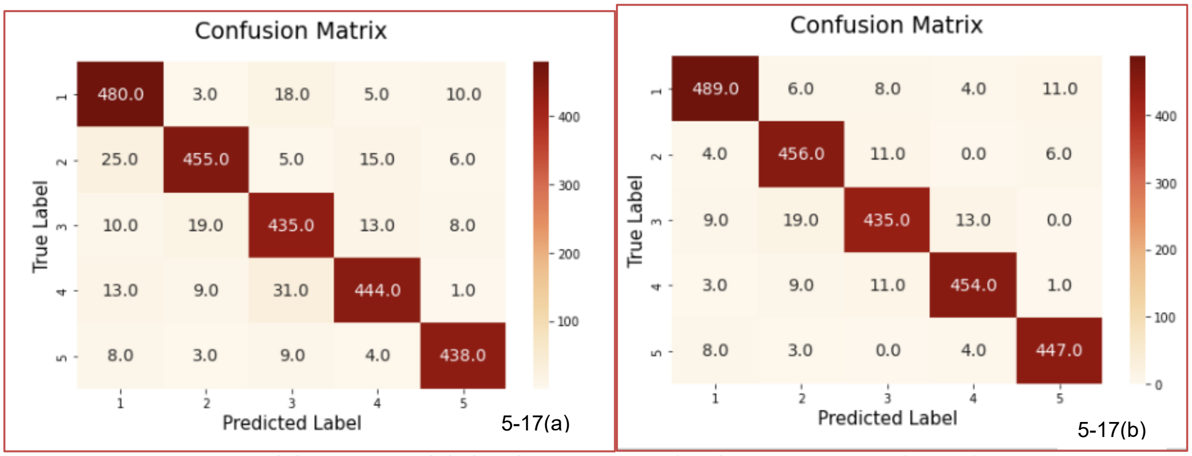
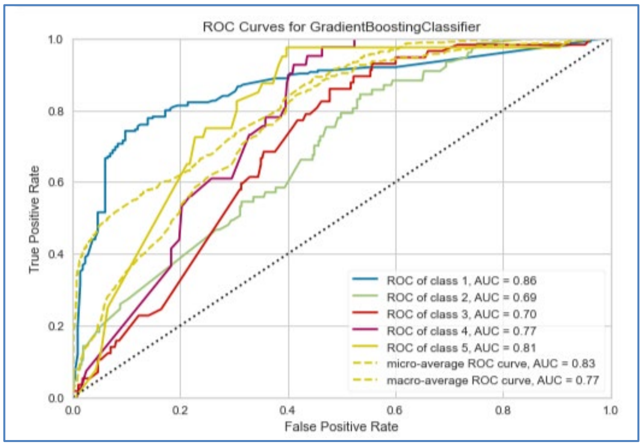
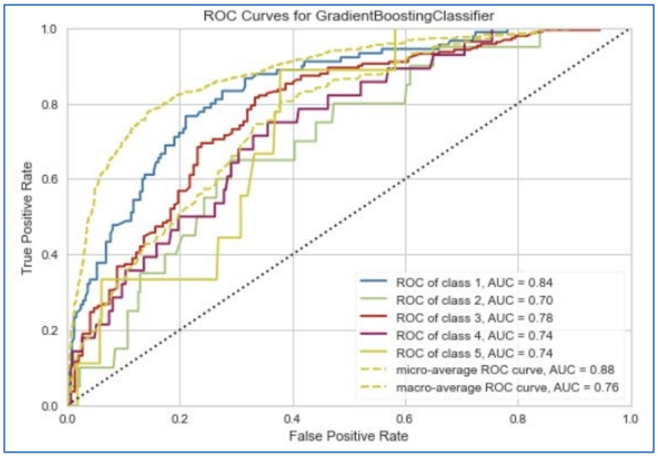


Figure 5-15 (a) and 5-17 (b) Confusion Matrix for Gradient Boost Classifier with Gridsearch CV and Randomizedsearch CV–Southeast



5-18 (a) ROC Curve for Tuned Gradient Boost Classifier using Gridsearch CV-Southeast



5-18(b) ROC Curve for Tuned Gradient Boost Classifier using Randomizedsearch CV-Southeast

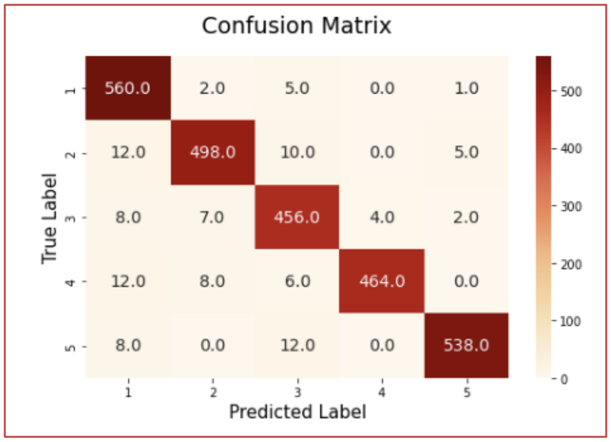
Table 5-14 Performance of Tuning Gradient Boost Classifier-Southeast

Performance	Accuracy	Macro F1-score
Gridsearch CV Training set	82.12%	86.12%
Gridsearch CV Test set	76.35%	76.73%
Randomizedsearch CV Training set	79.03%	79.15%
Randomizedsearch CV Test set	76.08%	76.20%

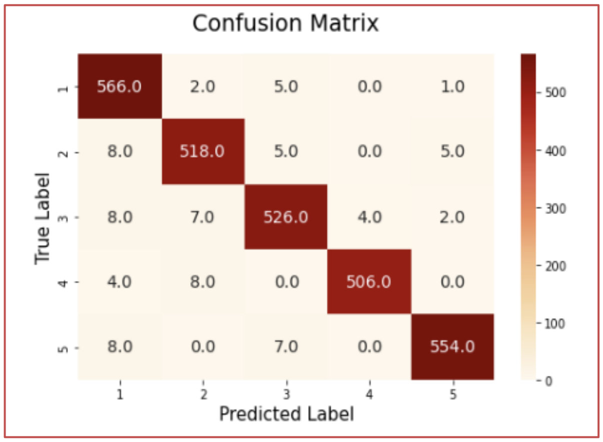
The ROC curve shown in Figure 5-18(a) has good AUC for the classes 1,4,5 that means it has good true positive rate (TPR) rate in those classes and macro-AUC for all classes of 0.77. Similarly, the ROC curve in Figure 5-18(b) shows a good AUC under classes 1 and 3 with macro-AUC for all classes of 0.76. The performance of gradient boost classifier for gridsearch and randomizedsearch are shown in Table 5-14, and it can be observed that macro F1-score of Gridsearch is 0.767 but the model is slightly overfitting. Nut in case of randomized CV, the macro F1-score is 0.762 with no overfitting.

5.2.1.2.10 Tuned XGBoost Classifier using Over-sampled Data

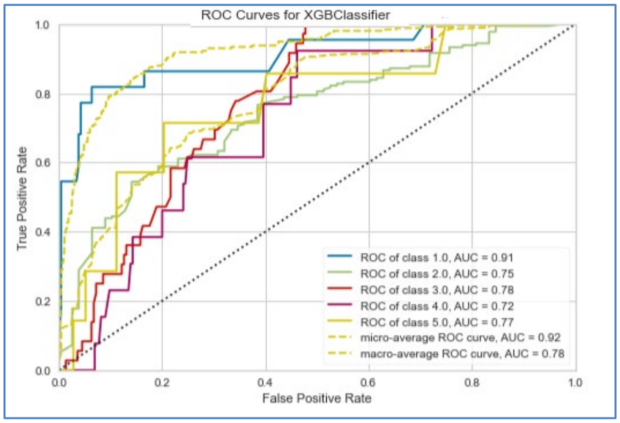
As the last model, XGBoost Classifier with hyperparameter tuning with respect to gridsearch CV and randomizedsearch CV was assessed in this study under the Northeast. The XGBoost method works similarly to the gradient boost classifier method but with much faster execution. The confusion matrix for both optimized search models and the ROC curves were developed and illustrated in Figure 5-19(a) and 5-19(b), Figure 5-20(a) and 5-20(b) respectively. The training set and test set performance for accuracy and F1-score are tabulated under Table 5-15. This model is not overfitting and has better scores than all the algorithms discussed above in this research. In the confusion matrix shown in Figure 5-19(a), it shows that there extremely low misclassifications under the class 1 and fewer under all the classes and the diagonal elements as has much higher values predicted rightly which indicates in improving the model performance. In the confusion matrix shown above, it shows that there are few misclassifications under all the classes and the diagonal elements as has much higher values predicted rightly compared to the confusion matrix in Figure 5-19(b) which indicates that model performance of randomizedsearch CV will better than the gridsearch CV.



5-19 (a) Confusion Matrix for XGBoost Classifier with Gridsearch CV-Southeast



5-19 (b) Confusion Matrix for XGBoost Classifier with Randomized search CV–Southeast



5-20 (a) ROC Curve for Tuning XGBoost Classifier
Gridsearch CV-Southeast

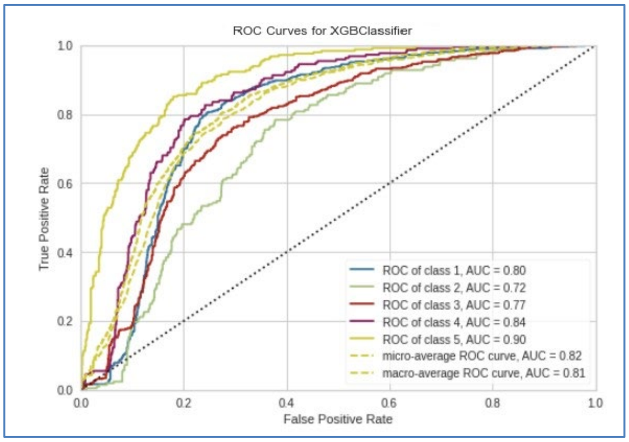


Figure 5-20 (b) ROC Curve for Tuning XGBoost Classifier
Randomized search CV-Southeast

The AUC for all the classes in gridsearch CV model has good score above 0.70 with macro average of 0.78 and the similarly, AUC for randomizedsearch CV has a macro average of 0.81 with classes 1 and 5 having maximum ROC. From the performance shown below, it is seen that randomizedsearch CV model has performed better with macro F1-score of 0.79.

Table 5-15 Performance of Tuning XGBoost Classifier-Southeast

Performance	Accuracy	Macro F1-score
Gridsearch CV Training Set	80.00%	80.02%
Gridsearch CV Test Set	77.14%	78.23%
Randomizedsearch CV Training Set	80.14%	79.17%
Randomizedsearch CV Test Set	78.50%	78.80%

5.2.1.3 Discussion on the Developed Models – Southeast

In this study, there are a total of 13 tree-based models that were developed to evaluate and compare the performance of each model to choose the best model to identify the critical variables in the Southeast. Figure 5-21 summarizes the test performance of all the models discussed in section 5.2.1.2.

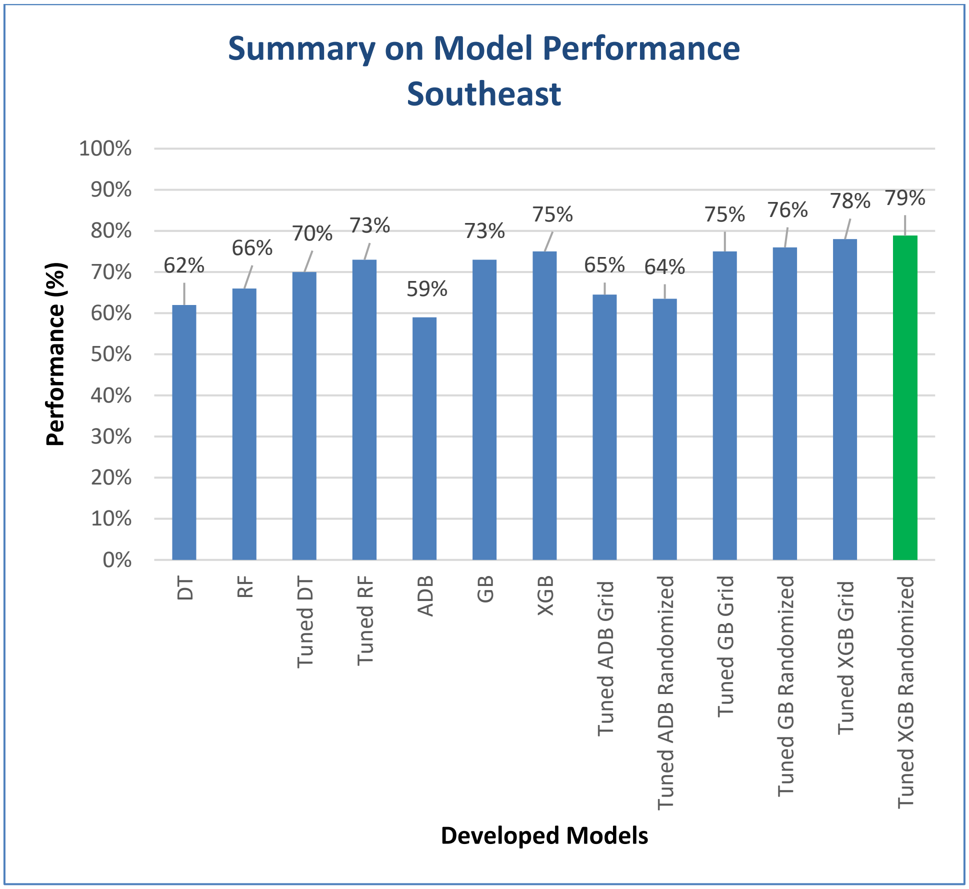


Figure 5-16 Summary on Models Test Performances-Southeast

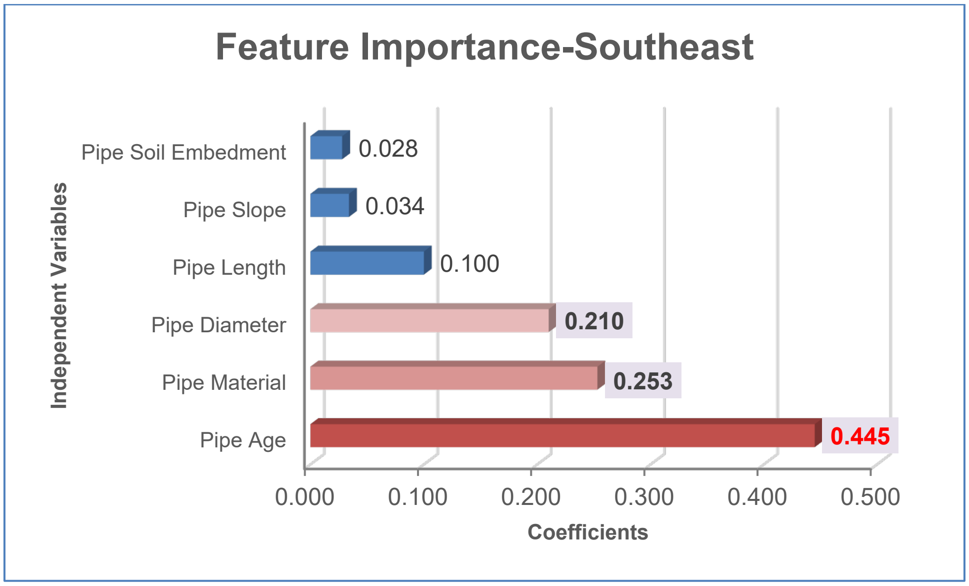


Figure 5-17 Feature Importance-Southeast

As the test performance of XGBoost based on Randomizedsearch CV algorithm is performed better than all the other models, the feature importance of the variables for this model was developed and is presented Figure 5-22. From this feature importance observation, we can conclude that pipe age is the high impact variable with 0.445 coefficient value for the PACP score. Pipe material and pipe diameter are the second most impacting variables on PACP score with 0.253 and 0.210 values, respectively. Pipe slope and pipe soil type has the least coefficient values and can be concluded that they have the least effect on the PACP score. In order to evaluate on this conclusion, XGBoost model was reconstructed to check its performance by including the pipe age, pipe material, pipe diameter and pipe length variables in the dataset and it was found that, the macro F1-score for this model was 0.746. Therefore, it can be decided that the critical variables mentioned above in this section can be considered to have more effect on the PACP score.

5.3.1 Southcentral

5.3.1.1 Dataset Summary

Table 5-16 presents the descriptive statistics of variables like pipe age, pipe diameter, pipe length and pipe slope for the dataset collected from the Southcentral US region.

Table 5-16 Descriptive Statistics of Variables in Southcentral.

	Pipe Age	Pipe Diameter	Pipe Length	Pipe Slope	PACP Score
count	4268	4268	4268	4268	4268
mean	40.65	11.26	269.00	0.66	1.96
std	22.39	9.97	223.00	1.18	1.33
min	1.00	4.00	0.50	-0.02	1.00
25%	19.00	6.00	114.10	0.04	1.00
50%	45.00	8.00	221.00	0.28	1.00
75%	52.00	10.00	371.00	0.60	3.00
max	121.00	96.00	2054.00	14.12	5.00

Pipe material considered included – VCP, PVC, UnReinCONC, RCP, DI, HDPE, CI, FRP. The full form of the pipe materials is presented in Appendix Section A. The soil type included in the dataset was found to be sand, silt, clay, loam, and rock. The dataset for developing the models was divided into 70%-train set and 30% test set. Table 5-17 shows the unique values of the independent variable of the Southcentral dataset.

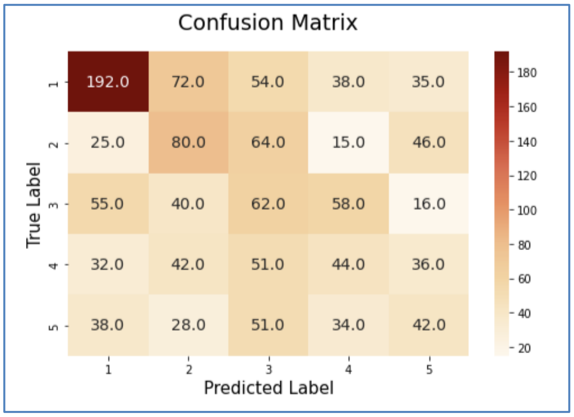
Table 5-17 Unique Values of Southcentral Dataset

Pipe Age	80
Pipe Diameter	32
Pipe Material	9
Pipe Length	3461
Pipe Slope	520
Pipe Soil	5
PACP Score	5

5.3.1.2 Results on Developed Models

5.3.1.2.1 Decision Tree (DT) Classifier with Imbalanced Dataset

The confusion matrix of test set developed for DT classifier method is shown in Figure 5-23. The last two columns have lower data points predicted in the class of 4 and 5 scores. This is by the fact that the dataset in that class 4 and 5 are low in number when compared to the class 1, 2, 3. The ROC curve for the DT classifier is plotted and presented in Figure 5-24, AUC for all classes except 1(which is 0.70) have low value below 0.60. This means that there is a low true prediction rate (TPR) in all 2,3,4, and 5 classes. The training set and test set performance is viewed in Table 5-18, and it is seen that the model displays poor accuracy and F1-score.



True \ Predicted	1	2	3	4	5
1	192.0	72.0	54.0	38.0	35.0
2	25.0	80.0	64.0	15.0	46.0
3	55.0	40.0	62.0	58.0	16.0
4	32.0	42.0	51.0	44.0	36.0
5	38.0	28.0	51.0	34.0	42.0

Figure 5-18 Confusion Matrix for Decision Tree Classifier–Southcentral

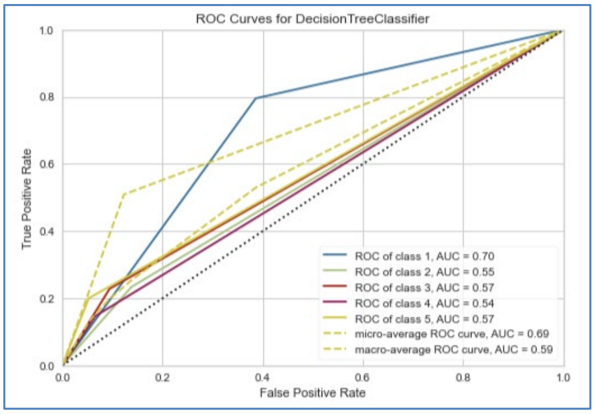


Figure 5-19 ROC Curve for Decision Tree Classifier-Southcentral

Table 5-18 Performance for Decision Tree Classifier-Southcentral

Performance	Accuracy	Macro F1-score
Training set	100.00%	100.00%
Test set	55.25%	56.25%

5.3.1.2.2 Random Forest (RF) Classifier with Imbalanced Data

The confusion matrix for RF classifier method is shown in Figure 5-25 below and can be inferred that, just like the DT classifier, the last two columns have less predicted values in the class of 4 and 5. The ROC curve for this model is plotted and shown in Figure 5-26, classes 1 and 3 have better ROC value than 2,4 and, 5 classes. The training set and test set performance is viewed in Table 5-5 and the accuracy and macro F1-score is better than DT classifier. But the model is not good as it is overfitting, and the performance is less than 70%.

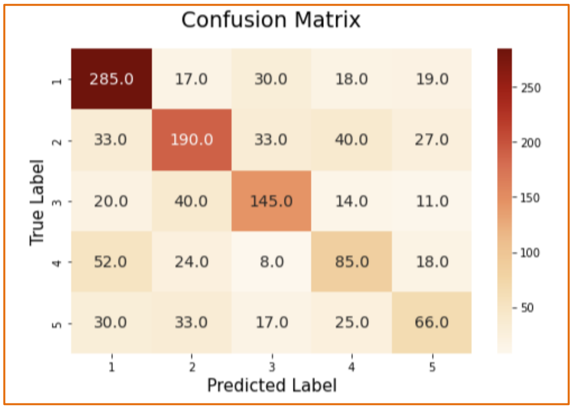


Figure 5-20 Confusion Matrix for Random Forest Classifier–Southcentral

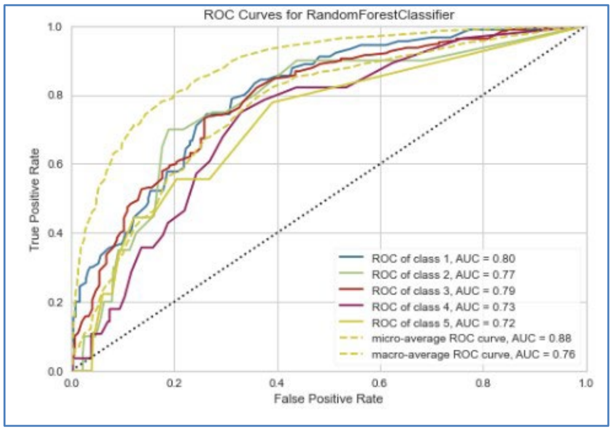


Figure 5-21 ROC Curve for Random Forest Classifier-Southcentral

Table 5-19 Performance of Random Forest Classifier-Southcentral

Performance	Accuracy	Macro F1-score
Training set	98.65%	98.25%
Test set	62.15%	61.52%

5.3.1.2.3 Tuned Decision Tree (DT) Classifier using Over-sampled Data

The grid parameters are this model are listed in Table 5-20. As the performance of DT model and RF model in the section 5.3.1.2.1 and 5.3.1.1.3. are not satisfactory, tuning of decision tree is carried out to optimize the given dataset. Confusion matrix for this method is given Figure 5-27 below. The ROC curve for the tuning DT Classifier is plotted and shown in Figure 5-28. The training set and test set performance is viewed in Table 5-21. After comparing the training set and testing set scores, the models are not overfitting but have poor performance.

Table 5-20 Tuned Decision Tree Classifier Grid Parameters

max_depth	35
max_leaf_nodes	10
min_samples_leaf	5
random_state	1

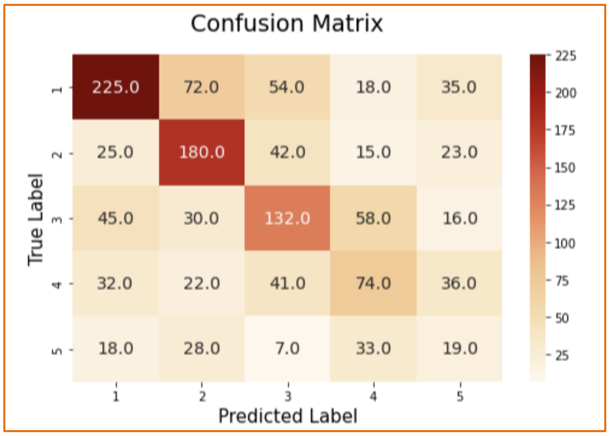


Figure 5-22 Confusion Matrix for Tuned Decision Tree Classifier-Southcentral

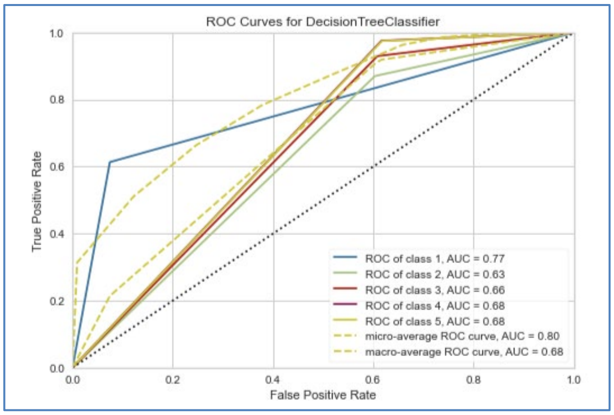


Figure 5-23 ROC Curve for Tuned Decision Tree Classifier-Southcentral

Table 5-21 Performance of Tuned Decision Tree Classifier – Southcentral

Performance	Accuracy	Macro F1-score
Training set	88.23%	88.12%
Test set	59.98%	59.75%

5.2.1.2.4 Tuned Random Forest (RF) Classifier using Oversampled Data

The grid parameters and confusion matrix for test set for this tuned RF model are shown in Table 5-22 and Figure 5-29 respectively. The values on diagonal element have better TPR than the model in section 5.2.1.2.2. The ROC curve for this model is plotted and shown in Figure 5-30. The training set and test set performance for accuracy and F1- score is viewed in Table 5-23. This model is not overfitting and has better performance than the RF model developed with default parameters.

Table 5-22 Tuning Random Forest Classifier Grid Parameters

Max_depth	10
max_samples	0.2
min_impurities_decrease	0.0001
N_estimates	150
random_state	1

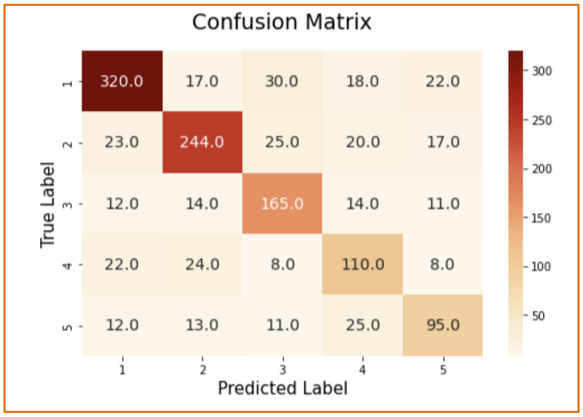


Figure 5-29 Confusion Matrix for Tuned Random Forest Classifier-Southcentral

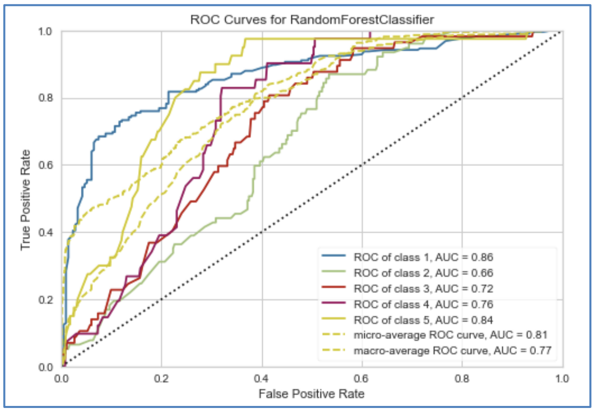


Figure 5-30 ROC Curve for Tuned Random Forest Classifier-Southcentral

Table 5-23 Performance of Tuned Random Forest Classifier-Southcentral

Performance	Accuracy	Macro F1-score
Training set	72.35%	73.56%
Test set	71.12%	72.22%

5.3.1.2.5 AdaBoost Classifier with Default Parameters using Imbalanced Data

AdaBoost Classifier with default parameter algorithm is one of the ensembles boosting algorithms employed in model development of this research. The confusion matrix for this model is shown in Figure 5-31. The actual predicted values distributed in all the classes are seen to have better in number when compared to RF model with default parameters. The ROC curve for the AdaBoost classifier is plotted as in Figure 5-32. The training set and test set performance for accuracy and F1- Score is viewed in Table 5-24. This model's performance is not satisfactory.

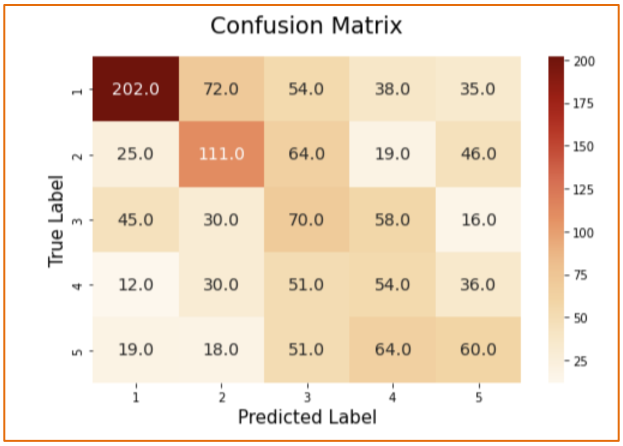


Figure 5-31 Confusion Matrix for AdaBoost Classifier–Southcentral

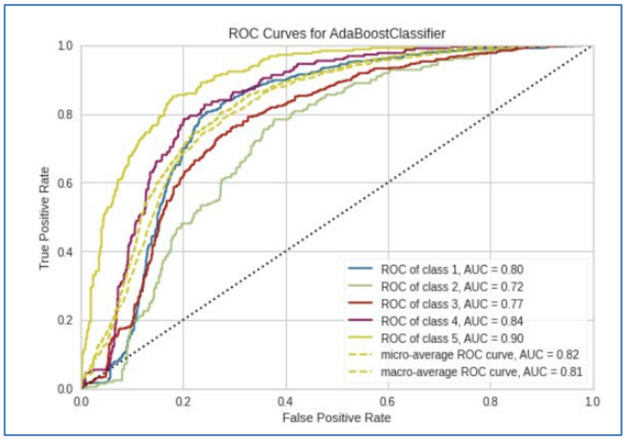


Figure 5-32 ROC Curve for AdaBoost Classifier–Southcentral

Table 5-24 Performance of AdaBoost Classifier – Southcentral

Performance	Accuracy	Macro F1-score
Training set	67.52%	67.65%
Test set	66.62%	65.9%

5.3.1.2.6 Gradient Boost Classifier with Default Parameters using Imbalanced Data

The Gradient boosting classifier with default parameter is computed and its confusion matrix is shown in Figure 5-33, and it is observed that the values on diagonal element are predicted rightly and the misclassification in all the classes are improved compared to AdaBoost model. The ROC curve for this model is plotted Figure 5-34 and the AUC for all classes have good score above 0.7 indicating that model has developed in good manner and will have improved performance. The training set and test set performance for accuracy and macro F1- score is viewed in Table 5-25. This model is having test macro F1-score of 73.22% and it is not overfitting and has better scores than AdaBoost Classifier

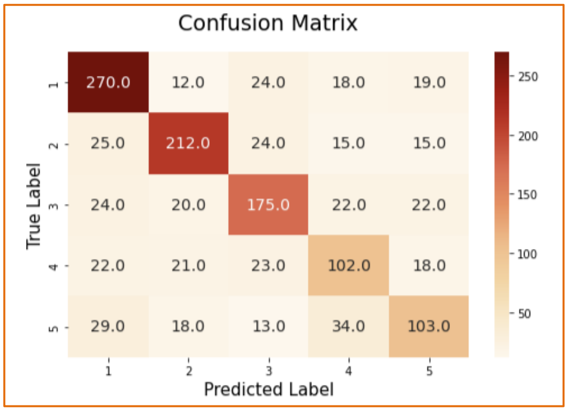


Figure 5-33 Confusion Matrix for Gradient Boost Classifier–Southcentral

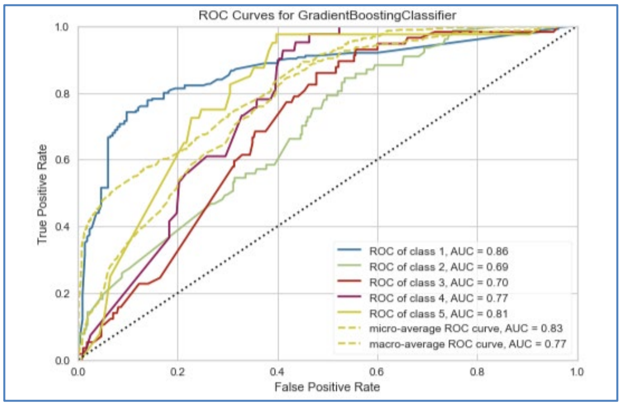


Figure 5-34 ROC Curve for Gradient Boost Classifier–Southcentral

Table 5-25 Performance of Gradient Boost Classifier-Southcentral

Performance	Accuracy	Macro F1-score
Training set	74.52%	74.68%
Test set	73.65%	73.22%

5.3.1.2.7 XGBoost Classifier with Default Parameters using Imbalanced Data

XGBoost is the other most effective algorithm in predicting classification. It computes better and faster than the other algorithms. The confusion matrix for this model is illustrated in Figure 5-35. The ROC curve for the XGBoost Classifier is developed and presented in Figure 5-36. The training set and test set performance for accuracy and F1- Score are listed in Table 5-26. This model is not overfitting and has better scores than other ensemble algorithms discussed above in this research. There is drastic increase in the predicted true values on diagonal element and misclassification in classes is low compared to the other tree- based models discussed earlier in this section.

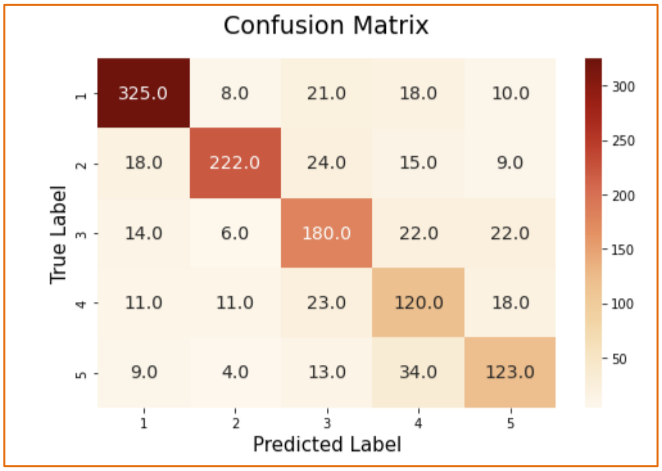


Figure 5-35 Confusion Matrix for XGBoost Classifier-Southcentral

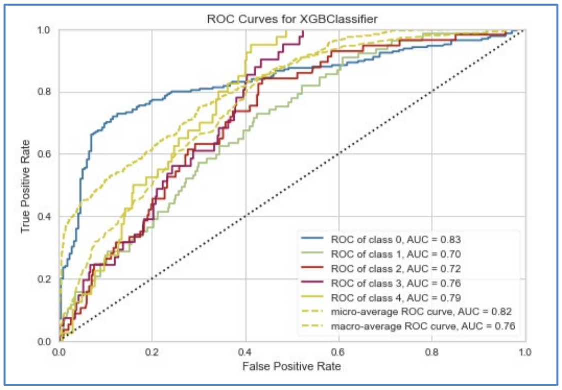


Figure 5-36 ROC Curve for XGBoost Classifier-Southcentral

Table 5-26 Performance of XGBoost Classifier-Southcentral

Performance	Accuracy	Macro F1-score
Training set	78.12%	78.88%
Test set	76.52%	76.75%

5.3.1.2.8 Tuned AdaBoost Classifier using Over-sampled Data

In this study, an approach of oversampling the dataset is carried out to treat the imbalanced data in the class of 4 and 5 PACP score by adopting SMOTE technique. Two ways of hyperparameter tuning were carried out namely gridsearch CV and randomizedsearch CV. The dataset was divided into categories of train set and validation set and the performance of model in validation set is considered important. Confusion matrix for test set for both gridsearch CV and randomizedsearch CV is shown in Figure 5-37(a) and Figure 5-37(b) respectively. The TPR values are improved and are found to be better than the model in section 5.3.1.2.5. The ROC curve is plotted in Figure 5-38(a) and 5-38(b) respectively and the AUC values for both models are improved and better compared to the model developed with imbalanced dataset in section 5.3.1.2.5. The training set and test set performance for accuracy and F1-score are listed in Table 5-27. This model is overfitting and has poor performance.

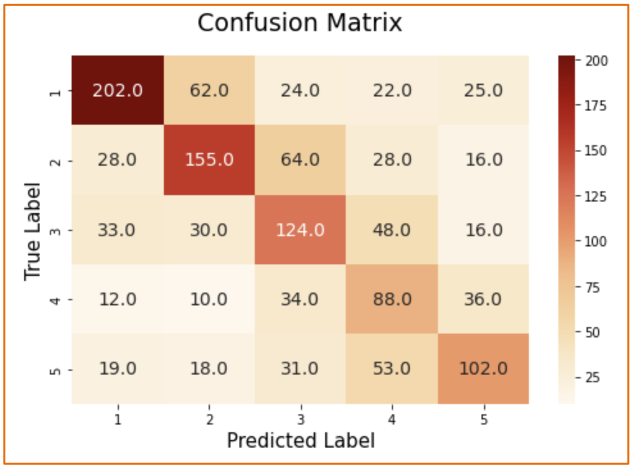


Figure 5-37(a) Confusion Matrix for Tuned AdaBoost Classifier with Gridsearch CV-Southcentral

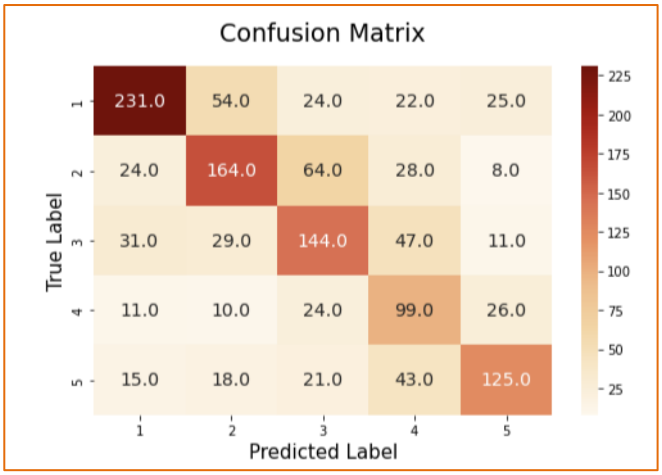


Figure 5-37(b) Confusion Matrix for Tuned AdaBoost Classifier with Randomizedsearch CV-Southcentral

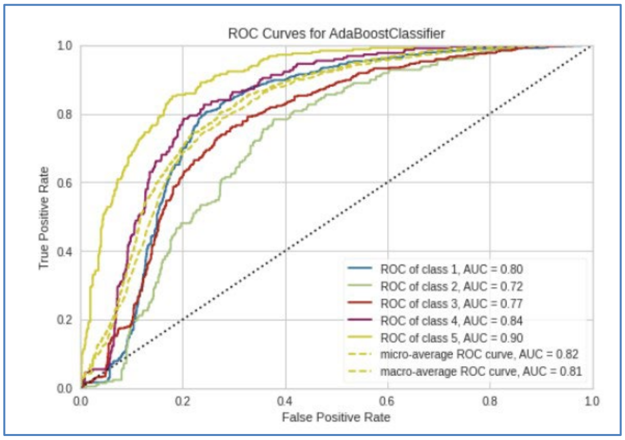


Figure 5-38(a) Tuned AdaBoost Classifier with Gridsearch CV-Southcentral

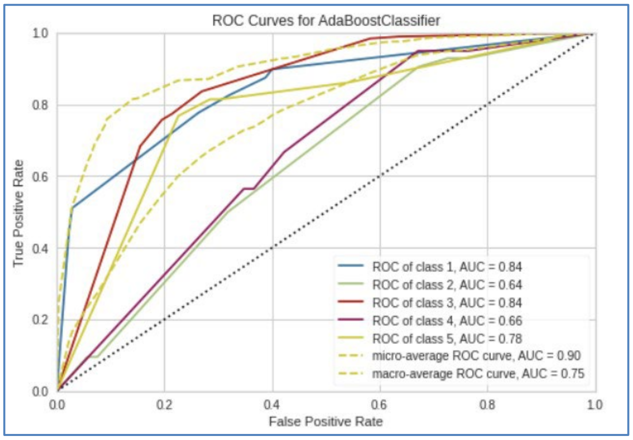


Figure 5-38(b) Tuned AdaBoost Classifier with Randomizedsearch CV-Southcentral

Table 5-27 Performance of Tuned AdaBoost Classifier-Southcentral

Performance	Accuracy	Macro F1-score
Gridsearch CV Training set	99.85%	99.18%
Gridsearch CV Test set	66.5%	66.85%
Randomizedsearch CV Training set	98.80%	98.62%
Randomizedsearch CV Test set	67.55%	67.25%

5.3.1.2.9 Tuned Gradient Boost Classifier using Over-sampled Data

The Gradient boosting algorithm was tuned with hyperparameters, and the confusion matrix is illustrated as presented in Figure 5-38(c) and Figure 5-38(d) for gridsearch CV and randomizedsearch CV, respectively. The ROC curve for these models is shown in Figures 5-39(a) and 5-39(b) below. Also, the performance of this model is tabulated in Table 5-28.

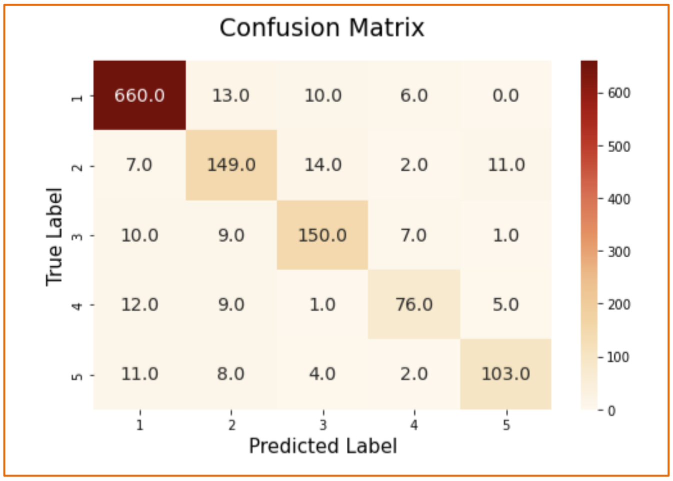


Figure 5-38(c) Confusion Matrix for Tuned Gradient Boost Classifier with Gridsearch CV-Southcentral

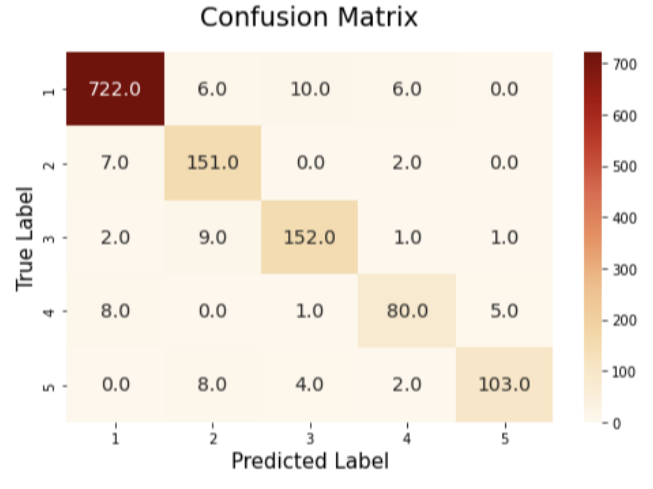


Figure 5-38(d) Confusion Matrix for Tuned Gradient Boost Classifier with Randomizedsearch CV-Southcentral

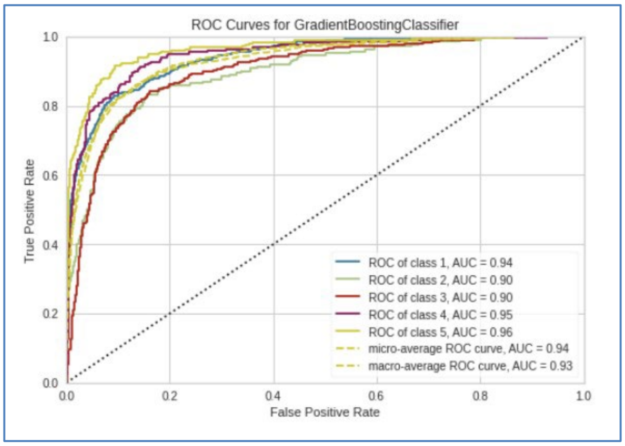


Figure 5-39(a) ROC Curve for Tuned Gradient Boost Classifier using Gridsearch CV-Southcentral

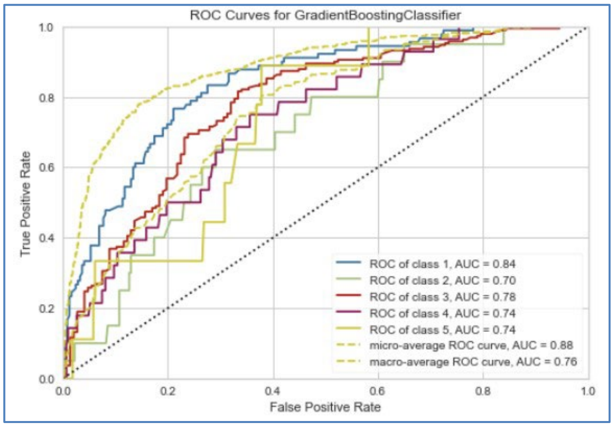


Figure 5-39(b) ROC Curve for Tuned Gradient Boost Classifier using Randomizedsearch CV-Southcentral

Table 5-28 Performance of Tuned Gradient Boost Classifier-Southcentral

Performance	Accuracy	Macro F1-score
Gridsearch CV Training set	80.55%	80.49%
Gridsearch CV Test set	74.99%	75.00%
Randomizedsearch CV Training set	82.12%	81.99%
Randomizedsearch CV Test set	76.52%	76.32%

5.3.1.2.10 Tuned XGBoost Classifier using Over-sampled Data

As the last model, XGBoost Classifier with hyperparameter tuning with respect to gridsearch CV and randomizedsearch CV assessed in this study under Southcentral. The Confusion matrix for both optimized search models are illustrated in Figure 5-40(a) and 5-40(b) and true predicted values on the diagonal element is improved on high rate and misclassification in all classes have reduced and resulted in good model performance. The ROC curves are plotted for both hyperparameter tuning models in Figure 5-41(a) and 5-41(b). The training set and test set performance for accuracy and macro F1-score than all other models discussed in Section 5.3.1. 2.

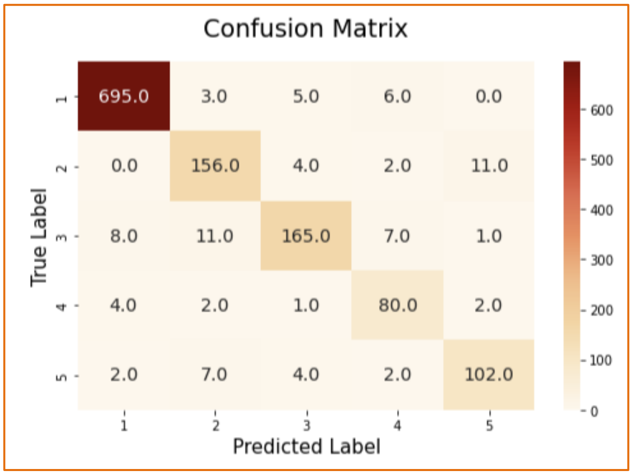


Figure 5-40(a) Confusion Matrix for Tuned XGBoost Classifier with Gridsearch CV-Southcentral

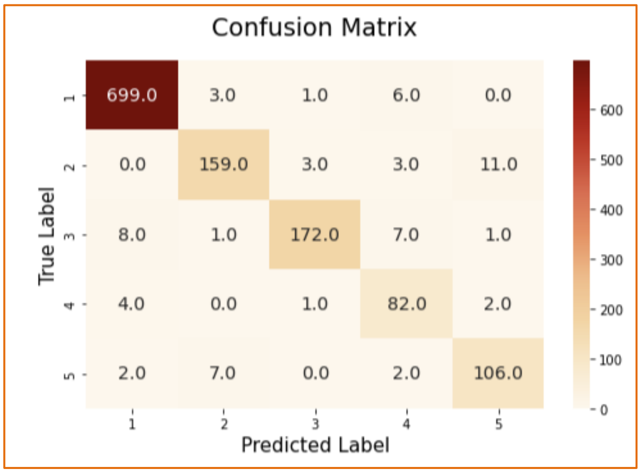


Figure 5-40(b) Confusion Matrix for Tuned XGBoost Classifier with Randomizedsearch CV-Southcentral

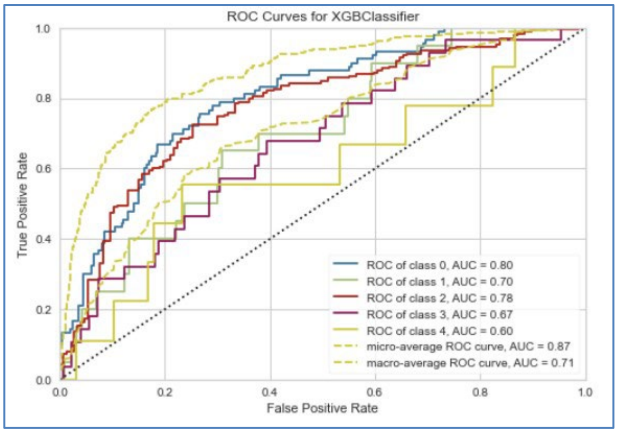


Figure 5-41(a) ROC curve for Tuned XGBoost Classifier Gridsearch CV-Southcentral

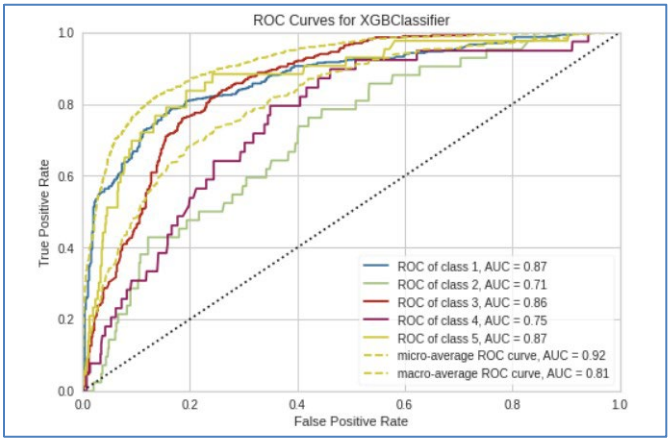


Figure 5-41(b) ROC curve for Tuned XGBoost Classifier Randomizedsearch CV-Southcentral

Table 5-29 Performance of XGBoost Classifier-Southcentral

Performance	Accuracy	Macro F1-score
Gridsearch CV Training set	80.23%	80.35%
Gridsearch CV Test set	77.65%	78.52%
Randomizedsearch CV Training set	79.12%	79.23%
Randomizedsearch CV Test set	78.86%	78.82%

5.3.1.3 Discussion on the Developed Models- Southcentral

In this study, there are a total of 13 tree-based models that were developed to evaluate and compare the performance of each model to choose the best model to identify the critical variables in the Southcentral. Figure 5-42 summarizes the test performance of all the models discussed in section 5.3.1.2.

As the test performance of XGBoost based on randomizedsearch CV algorithm is performed better than all the other models, the feature importance of the variables for this model was developed and is presented Figure 5-43. It can be seen that pipe material and pipe age exhibit the high impact on the PACP score with the relative importance of 0.36 and 0.37 and pipe diameter is the second most impacting variable with the co-efficient of 0.20. The pipe length and slope have low scores of 0.035 and 0.028. Finally pipe soil type has the lowest score of 0.0015 and its shows no impact on the PACP score. Inorder to evaluate on this conclusion, XGBoost model was reconstructed to check its performance by including the pipe age, pipe material, pipe diameter and pipe length variables in the dataset and it was found that, the macro F1-score for this model was 0.80 and this model has outperformed than the previous developed model with all the independent variables. Therefore, it can be decided that the critical variables mentioned above in this section can be considered to have more effect on the PACP score by neglecting pipe slope and pipe native soil type.

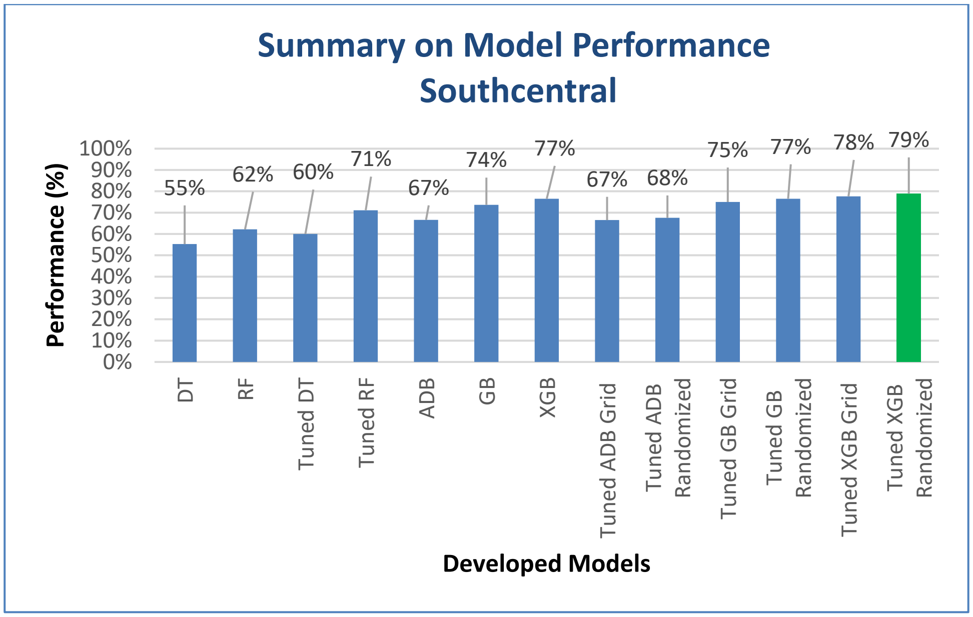


Figure 5-42 Summary on Models Test Performance-Southcentral

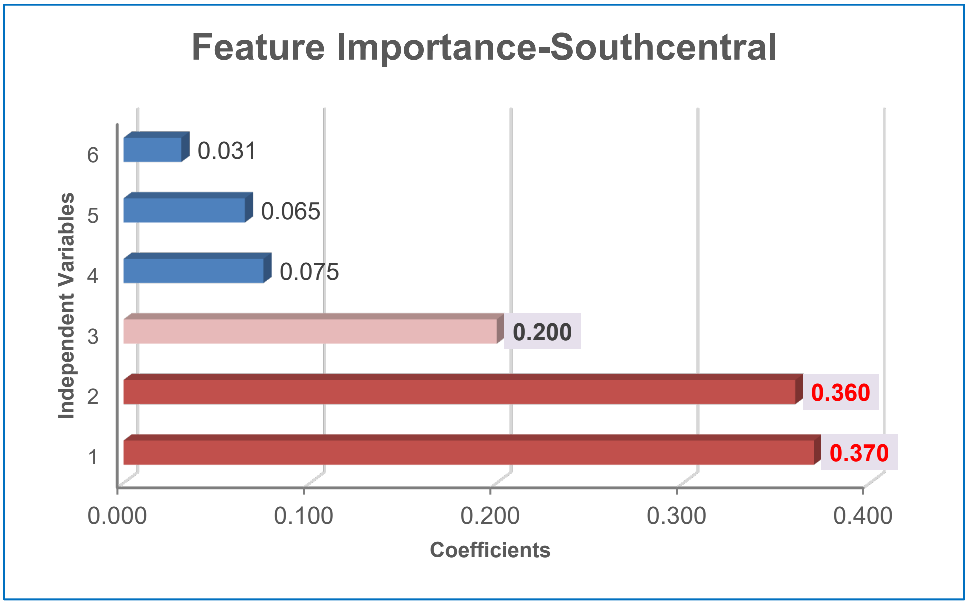


Figure 5-43 Feature importance of Independent Variables for Southcentral

5.4.1 Midwest

5.4.1.1 Dataset Summary

Table 5-30 presents the descriptive statistics of summary of variables like pipe age, pipe diameter, pipe length and pipe slope for the dataset collected from Midwest US region.

Table 5-30 Descriptive Statistics of Variables in Midwest

	Pipe Age	Pipe Diameter	Pipe Length	Pipe Slope	PACP Score
count	2088	2088	2088	2088	2088
mean	53.75	22.33	218.93	0.12	2.43
std	19.00	14.33	148.74	0.17	0.96
min	1.00	6.00	1.00	0.00	1.00
25%	51.00	10.00	96.00	0.00	2.00
50%	53.00	18.00	197.00	0.01	2.00
75%	63.00	33.00	317.00	0.23	3.00
max	107.00	96.00	1238.00	0.60	5.00

Pipe material considered included – VCP, PVC, RCP, AC, PCCP, RCP, DI, CI, FRP, others, CMP. The full form of the pipe materials is presented in Appendix Section A. The native soil type included in the dataset was found to be sand, sand fines, silt, clay, and loam. The dataset for developing the models was divided into 70%-

train set and 30% test set. Table 5-31 shows the unique values of independent variables of the Midwest dataset.

Table 5-31 Unique values of Midwest Dataset

Pipe Age	59
Pipe Diameter	24
Pipe Material	11
Pipe Length	521
Pipe Slope	1013
Pipe Native Soil	5
PACP Score	5

5.4.1.2 Results on Developed Models

5.4.1.2.1 Decision Tree (DT) Classifier with Imbalanced Dataset

The confusion metric developed for DT model with default parameter is visualized in Figure 5-44, it can be seen the last two columns have lower data points predicted in the class of 4 and 5 scores. This is by the fact that the dataset in that class 4 and 5 are low in number when compared to the class 1, 2, 3. The ROC curve for the DT model is plotted and shown in Figure 5-45, and it shows that the class1 has good ROC score above 0.70 and rest all classes have poor AUC value, this is of the fact that there are high misclassifications in those classes. The training set and test set performance is viewed in Table 5-32, and it is seen that the model displays poor accuracy and macro F1- score of 58.77% and 56.88% respectively and the model is overfitting.

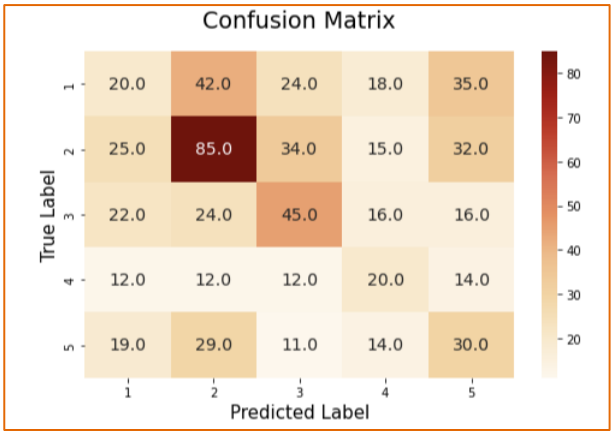


Figure 5-44 Confusion Matrix for Decision Tree Classifier–Midwest

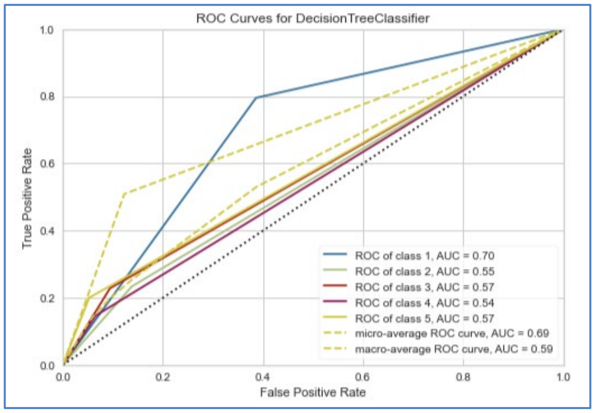


Figure 5-45 ROC Curve for Decision tree Classifier-Midwest

Table 5-32 Performance of Decision Tree -Midwest

Performance	Accuracy	Macro F1-score
Training set	99.72%	99.72%
Test set	58.77%	56.88%

5.4.1.2.2 Random Forest (RF) Classifier Method with Imbalanced Data

The RF classifier confusion matrix for test set is illustrated in Figure 5-46 below and the ROC curve for this model plotted and presented as in Figure 5-47 and can be seen that, similar to DT classifier, the last two columns have less predicted values in the class of 4 and 5.The ROC curve for RF classifier with default parameter is shown in Figure 5-47 and it is observed that all classes have the low AUC value due to high misclassification. The performance of training and test data can be reviewed in is viewed in Table 5-33. This model is shown to be overfitting and has performed poorly with accuracy and Macro F1-score less than 70%.

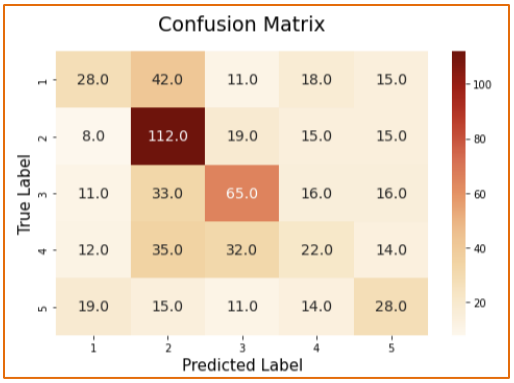


Figure 5-46 Confusion Matrix Random Forest Classifier-Midwest

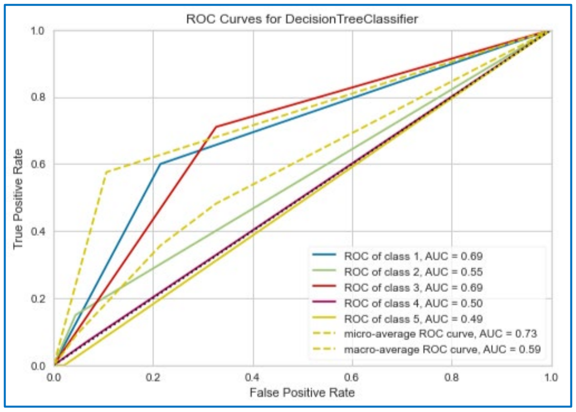


Figure 5-47 ROC Curve Random Forest Classifier-Midwest

Table 5-33 Performance of Random Forest Classifier–Midwest

Performance	Accuracy	Macro F1-score
Training set	99.86%	99.86%
Test set	53.74%	54.71%

5.4.1.2.3 Tuned Decision Tree (DT) Classifier using Over-sampled Data

Table 5-34 shows the grid parameters for tuned DT Classifier model. As discussed in earlier sections 5.4.1.2.1 and 5.4.1.1.2 the performance of the models is not satisfactory, so tuning technique employed to optimize the given dataset to achieve good performance.

Table 5-34 Tuned Decision Tree Classifier Grid Parameters

max_depth	35
max_leaf_nodes	15
min_samples_leaf	5
random_state	1

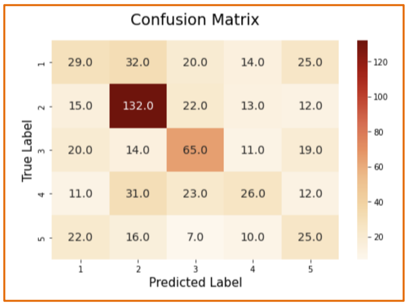


Figure 5-47(a) Confusion Matrix for Tuned Decision Tree Classifier-Midwest

The confusion matrix of test set for this model is presented in Figure 5-47(a), and it can be observed that the values on diagonal element has increased and the misclassification in the classes of 4and 5 has decreased in number compared to the confusion matrix developed in the section 5.4.1.2.1. In addition to this, The ROC curve for the tuning DT Classifier is plotted and shown in Figure 5-48, and all the classes have low AUC values below 0.70. The training set and test set performance are viewed in Table 5-35. The performance of the tuned DT classifier shows that model is overfitting and has a poor F1-score of 53.05%.

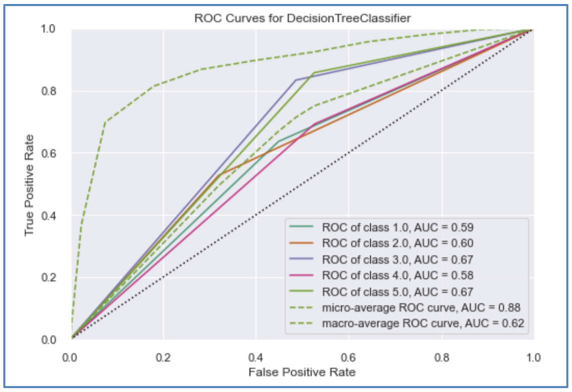


Figure 5-48 ROC Curve for Tuned Decision Tree Classifier-Midwest

Table 5-35 Performance of Tuned Decision Tree Classifier-Midwest

Performance	Accuracy	Macro F1-score
Training set	58.86%	66.32%
Test set	53.41%	53.05%

5.4.1.2.4 Tuned Random Forest (RF) Classifier using Oversampled Data

The confusion matrix for the tuned RF method is shown in Figure 5-49. And the grid parameters for building this RF model are shown in Table 5-36. The ROC curve for this model is plotted and shown in Figure 5-50. The training set and test set performance for accuracy and F1- Score are viewed in Table 5-37. This model is not overfitting.

Table 5-36 Tuning Random Forest Grid Parameters

max_depth	10
max_samples	0.2
min_impurities_decrease	0.0001
N_estimates	150
random_state	1

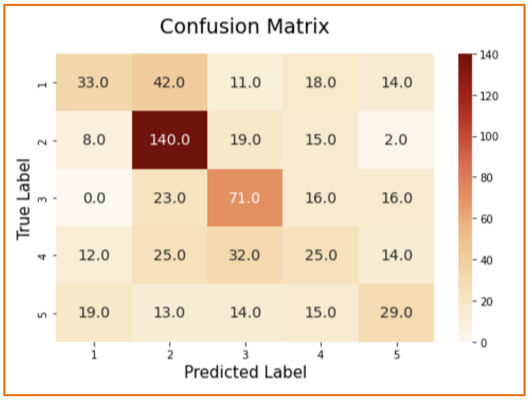


Figure 5-49 Confusion Matrix for Tuned Random Forest Classifier-Midwest

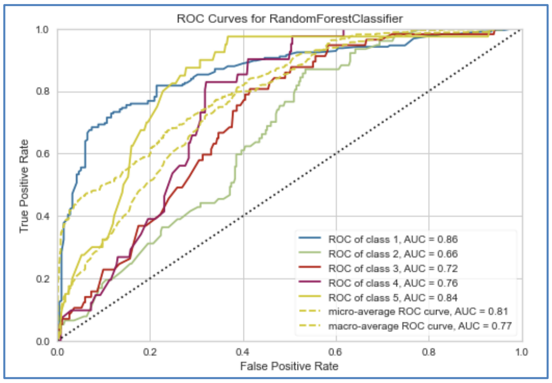


Figure 5-50 ROC Curve for Tuned Random Forest Classifier-Midwest

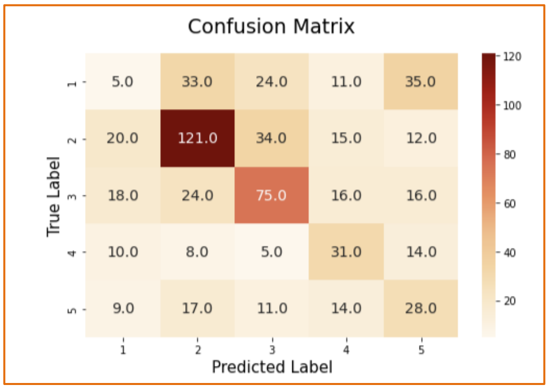
The TPR values are increased, compared to the model shown in section 5.4.1.2.2. Also, the misclassification in classes 1, 2 3 has decreased resulting in better performance. The AUC values in the figure 5-50 have improved and shows a good score of 0.77.

Table 5-37 Performance of Tuned Random Forest Classifier-Midwest

Performance	Accuracy	Macro F1-score
Training set	70.12%	70.35%
Test set	68.02%	69.17%

5.4.1.2.5 AdaBoost Classifier with Default Parameters using Imbalanced Data

The confusion matrix for the AdaBoost classifier with default parameters is shown in Figure 5-51 below and it has a poor predicting value with that of the true values. The ROC curve for the AdaBoost classifier is plotted and shown in Figure 5-52. The training set and test set performance for accuracy and F1- score are viewed in Table 5-38. This model is not overfitting.



	1	2	3	4	5
1	5.0	33.0	24.0	11.0	35.0
2	20.0	121.0	34.0	15.0	12.0
3	18.0	24.0	75.0	16.0	16.0
4	10.0	8.0	5.0	31.0	14.0
5	9.0	17.0	11.0	14.0	28.0

Figure 5-51 Confusion Matrix for Adaboost Classifier-Midwest

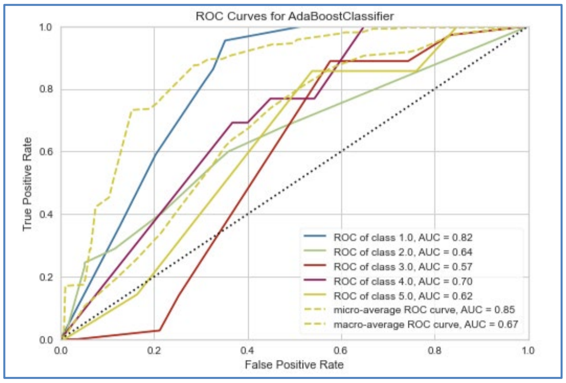


Figure 5-52 ROC Curve for AdaBoost Classifier-Midwest

Table 5-38 Performance of AdaBoost Classifier-Midwest

Performance	Accuracy	Macro F1-score
Training set	55.25%	53.44%
Test set	54.28%	52.99%

5.4.1.2.6 Gradient Boost Classifier with Default Parameters using Imbalanced Data

The gradient boosting classifier with default parameters is computed, and its confusion matrix is shown in Figure 5-53 below. The ROC curve for the gradient classifier is plotted and shown in Figure 5-54. The training set and test set performance for accuracy and F1- score is viewed in Table 5-39. This model is not overfitting and has better scores than AdaBoost classifier.

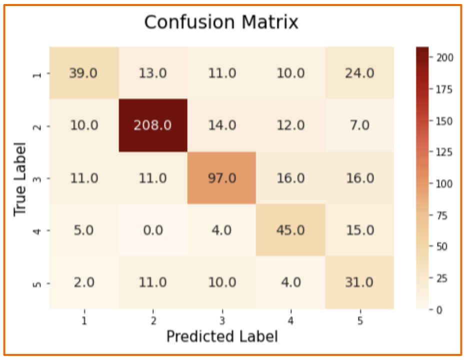


Figure 5-53 Confusion Matrix for Gradient Boost Classifier–Midwest

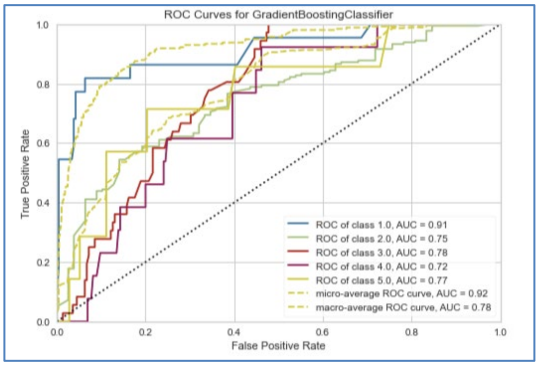


Figure 5-54 ROC Curve for Gradient Boost Classifier–Midwest

Table 5-39 Performance of Gradient Boost Classifier-Midwest

Performance	Accuracy	Macro F1-score
Training set	65.66%	65.89%
Test set	64.52%	64.00%

5.4.1.2.7 XGBoost Classifier with Default Parameters using Imbalanced Data

XGBoost is the other most effective algorithm in predicting classification. It computes better and faster than the other algorithms. The confusion matrix and ROC curve for the XGBoost Classifier are developed and presented in Figure 5-55 and Figure 5-56, respectively. The training set and test set performance for accuracy and F1- score are listed in Table 5-40. This model is not overfitting and has better scores than other ensemble algorithms discussed above in this research.

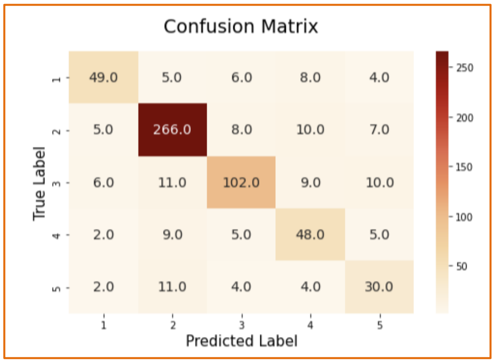


Figure 5-55 Confusion Matrix for XGBoost Classifier–Midwest

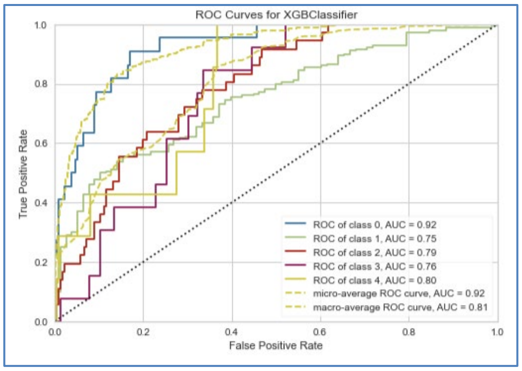


Figure 5-56 ROC Curve for XGBoost Classifier-Midwest

Table 5-40 Performance of XGBoost Classifier-Midwest

Performance	Accuracy	Macro F1-score
Training set	76.12%	76.88%
Test set	75.22%	75.15%

5.4.1.2.8 Tuned AdaBoost Classifier using Over-sampled Data

A tuned Adaboost classifier was developed with two hyperparameters tuning methods, namely gridsearch CV and randomizedsearch CV, to optimize the model performance. The confusion matrix for those two models is presented in Figure 5-57(a) and 5-57(b) below. The ROC curve for this both gridsearch CV and randomizedsearch CV is shown in the model is illustrated in Figure 5-58(a) and 5-58(b). The confusion matrix and the ROC curve understanding for this section is like the explanation in Section 5.3.1.2.8 The training set and test set performance for accuracy and F1- score are listed in Table 5-41. This model is overfitting and has poor performance.

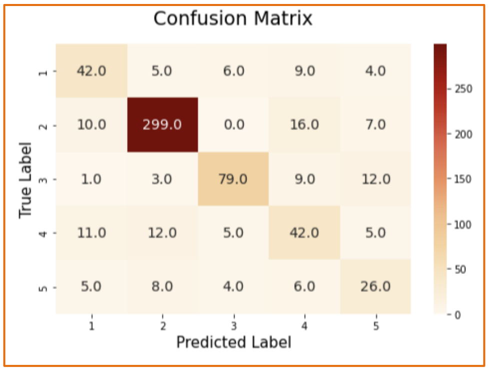


Figure 5-57(a) Confusion Matrix for Tuned AdaBoost Classifier with Gridsearch CV–Midwest

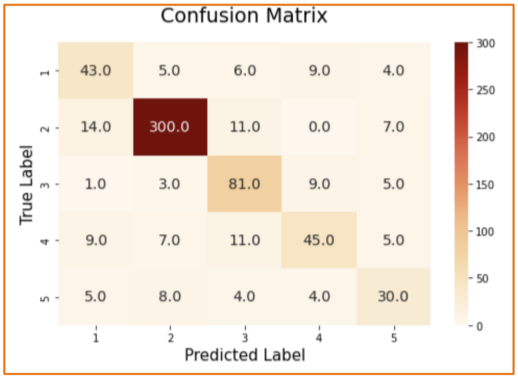


Figure 5-57(b) Confusion Matrix for Tuned AdaBoost Classifier with Randomizedsearch CV–Midwest

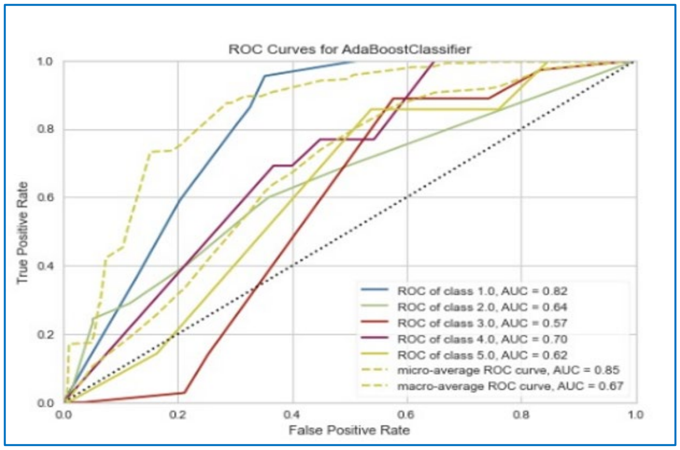


Figure 5-58(a) ROC Curve for Tuned AdaBoost Classifier with Gridsearch CV-Midwest

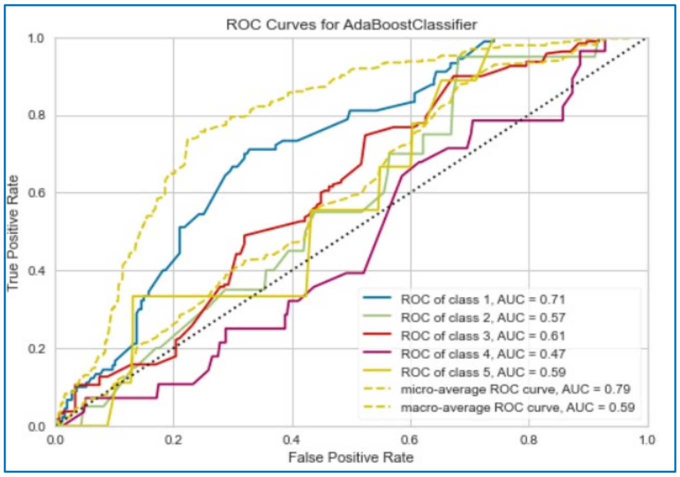


Figure 5-58(b) ROC Curve for Tuned AdaBoost Classifier with Randomizedsearch CV-Midwest

Table 5-41 Performance of Tuned AdaBoost Classifier-Midwest

Performance	Accuracy	Macro F1-score
Gridsearch CV Training set	99.74%	99.74%
Gridsearch CV Test set	66.5%	66.85%
Randomizedsearch CV Training set	98.80%	98.62%
Randomizedsearch CV Test set	69.55%	69.25%

5.4.1.2.9 Tuned Gradient Boost Classifier using Over-sampled Data

The gradient boosting algorithm was tuned with hyperparameters with gridsearch CV and randomizedsearch CV, and the confusion matrix for those models is illustrated in Figures 5.59(a) and 5.59(b). The predicted value on the diagonal area is evidence that the model performance is rightly evaluated. For the 4 and 5 classes wrongly, predicted values are accounting to 0 which is a good, developed model. The ROC curve for those models is plotted in Figures 5-60(a) and 5-60(b). Also, Performance is tabulated in Table 5-42.

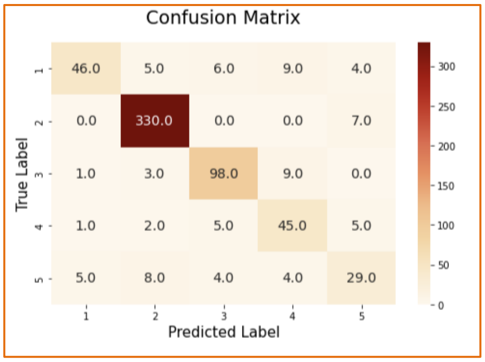


Figure 5-59(a) Confusion Matrix for Gradient Boost Classifier with Gridsearch CV–Midwest

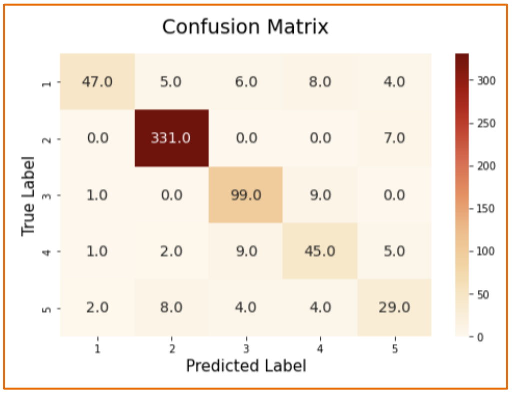


Figure 5-59(b) Confusion Matrix for Gradient Boost Classifier with Randomizedsearch CV–Midwest

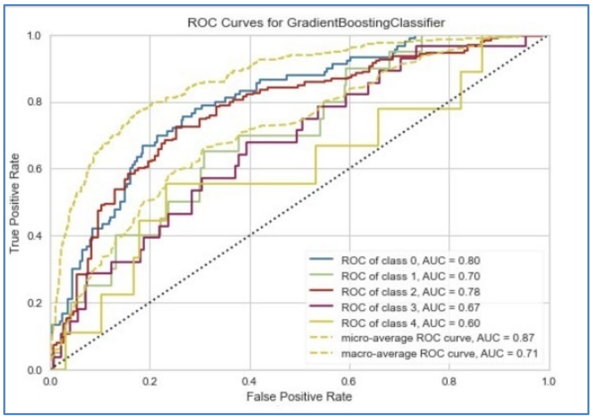


Figure 5-60(a) ROC curve for Tuned Gradient Boost Classifier for Gridsearch CV-Midwest

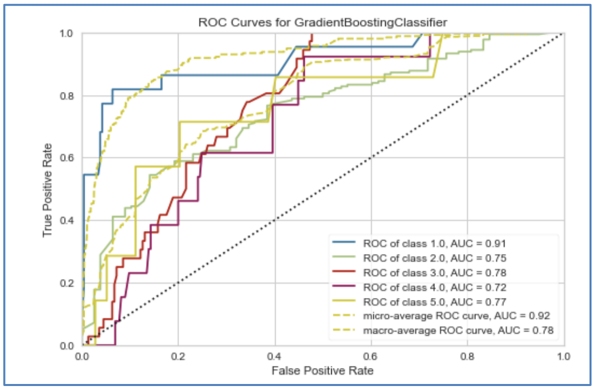


Figure 5-60(b) ROC curve for Tuned Gradient Boost Classifier
for Randomizedsearch CV-Midwest

The AUC for gridsearch CV has an overall good score for all classes above 0.70 with macro average of 0.78, which is good indication that the values are predicted rightly. Similarly, AUC for randomizedsearch CV has a macro average of 0.78, which shows that randomizedsearch CV is the better model than gridsearch CV.

Table 5-42 Performance of Tuning Gradient Boost Classifier-Midwest

Performance	Accuracy	Macro F1-score
Gridsearch CV Training set	85.25%	85.25%
Gridsearch CV Test set	73.20%	73.60%
Randomizedsearch CV Training set	81.12%	80.99%
Randomizedsearch CV Test set	76.33%	76.32%

5.4.1.2.10 Tuned XGBoost Classifier using Over-sampled Data

As the last model, XGBoost Classifier with hyperparameter tuning with respect to gridsearch CV and randomizedsearch CV was assessed in this study under the Midwest. The XGBoost method works similarly to the gradient boost classifier method but with much faster execution. The confusion matrix for both optimized search models and the ROC curves were developed and illustrated in Figures 5-61(a),5-61(b), 5-62(a), and 5-62(b). The training set and test set performance for accuracy and F1- score are tabulated under Table 5-43. This model is not overfitting and has better scores than all the algorithms discussed above in this research.

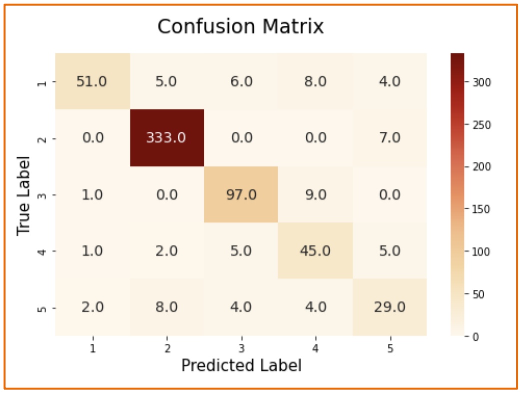


Figure 5-61(a) Confusion Matrix for XGBoost Classifier with Gridsearch CV–Midwest

In the confusion matrix shown above, it shows that there are few misclassifications under all the classes and the diagonal elements as has much higher values predicted rightly which indicates in improving the model performances.

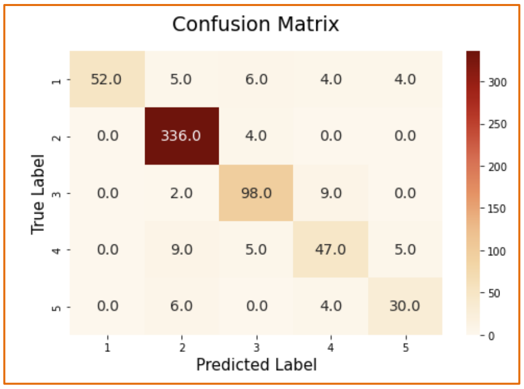
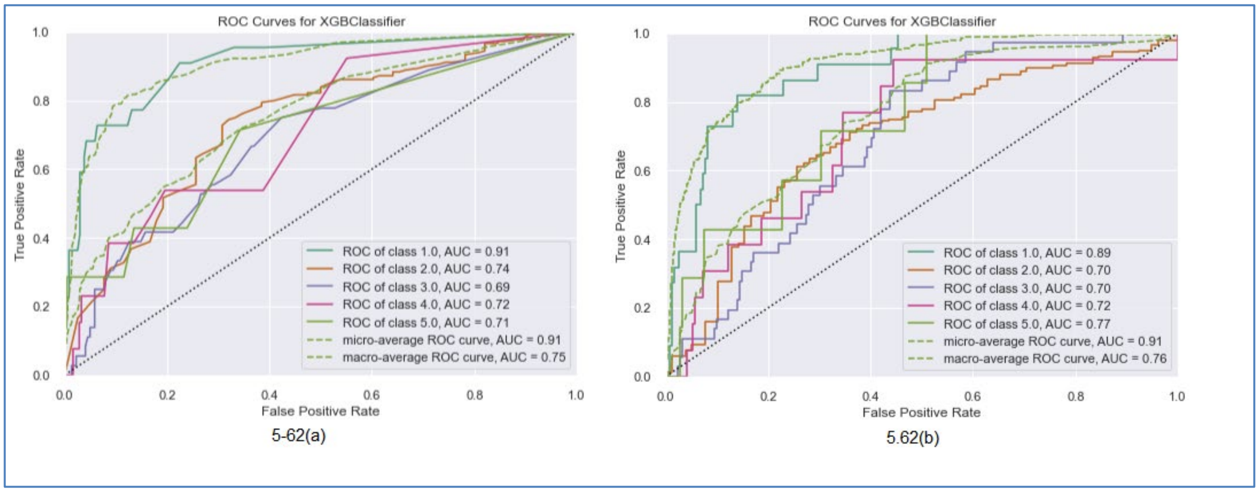


Figure 5-61(b) Confusion Matrix for XGBoost Classifier with Randomizedsearch CV–Midwest

In the confusion matrix shown above, it shows that there are few misclassifications under all the classes and the diagonal elements as has much higher values predicted rightly compared to the confusion matrix in Figure 5-61(a) which indicates that model performance of randomizedsearch CV will better than the gridsearch CV.



5-62(a)

5.62(b)

Figure 5-62(a) and 5.62(b) ROC Curve for Tuned XGBoost Classifier with Gridsearch CV and Randomizedsearch CV-Midwest

The AUC for gridsearch CV has an overall good score for all classes above and near to 70% with macro average of 75%, which is good indication that the values are predicted rightly. Similarly, AUC for randomizedsearch CV has a macro average of 78% which shows that randomizedsearch CV is the better model than gridsearch CV.

Table 5-43 Performance of XGBoost Classifier-Midwest

Performance	Accuracy	Macro F1-score
Gridsearch CV Training set	82.23%	83.35%
Gridsearch CV Test set	73.65%	75.52%
Randomizedsearch CV Training set	77.12%	77.32%
Randomizedsearch CV Test set	76.98%	76.88%

5.4.1.3 Discussion on the Developed Models-Midwest

In this study, as mentioned earlier, there are a total of 13 different models that were developed to evaluate and compare the performance of each algorithm to select the best model to prioritize the inspection in the Midwest. Figure 5-63 summarizes the test performance of all the models discussed in section 5.4.1.2. And Figure 5-64 visualizes the feature importance plotted for the best selected model.

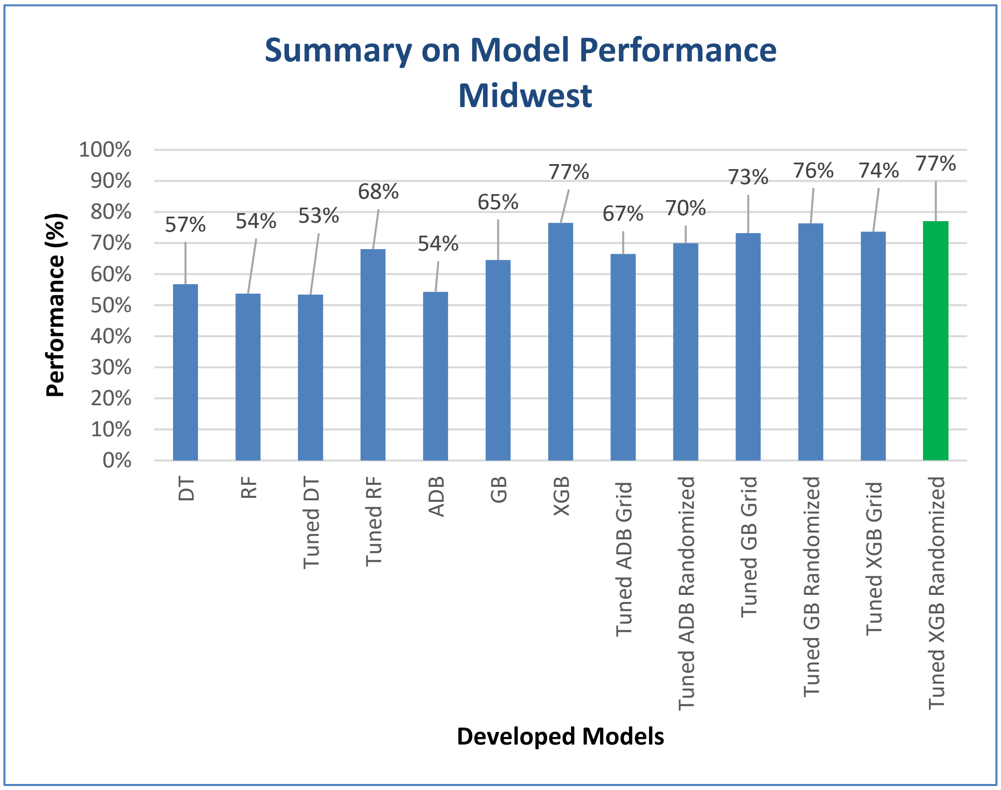


Figure 5-63 Summary on Models Test Performance-Midwest

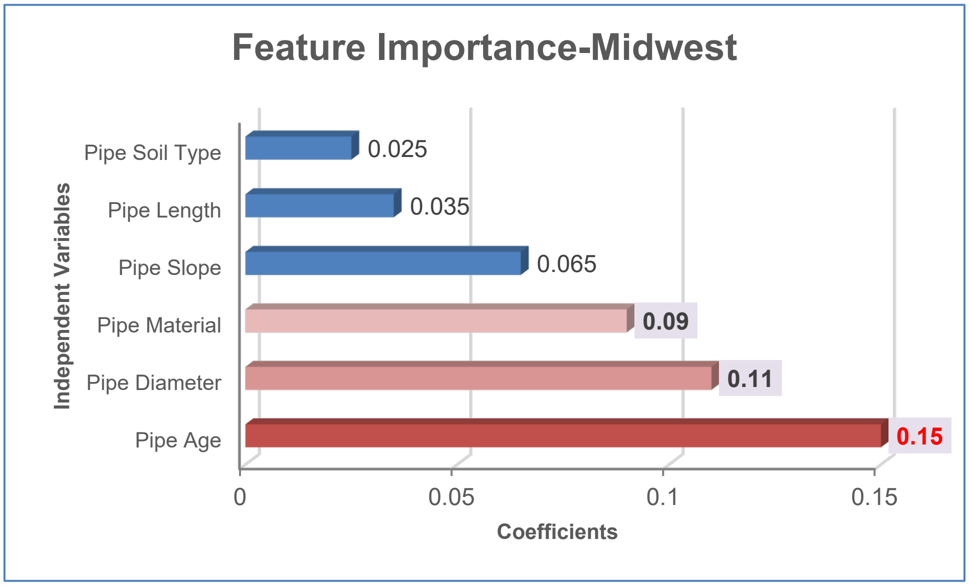


Figure 5-64 Feature Importance-Midwest

Figure 5-63 depicts the test performance of the XGBoost randomized search algorithm showing the best performance of all the other models developed for the Midwest. The feature importance of the variables for the XGBoost model is generated and is illustrated in Figure 5-64. It can be seen that pipe age (0.15), pipe diameter (0.11), and pipe material (0.09) are the top three critical variables affecting the PACP score. The pipe slope (0.065) in this region shows a more significant effect on the PACP score. Pipe length and pipe native soil type have low scores of 0.035 and 0.025. Inorder to evaluate on this conclusion, XGBoost model was reconstructed to check its performance by including the pipe age, pipe material, pipe diameter and pipe slope variables in the dataset and it was found that, the macro F1-score for this model was 0.691 and this model has slightly lower performance than the previous developed model with all the independent variables. Therefore, it can be decided that the identified critical variables for the Miswest region needs more inspection data to finalise on the critical variables.

5.5.1 Northeast

5.5.1.1 Dataset Summary

Table 5-44 summarizes descriptive statistics on variables like pipe age, pipe diameter, pipe length, and pipe slope for the dataset collected from Northeast.

Table 5-44 Descriptive Statistics of Variables in the Northeast

	Pipe Age	Pipe Diameter	Pipe Length	Pipe Slope	PACP Score
count	4869	4869	4869	4869	4869
mean	33.16	8.15	166.01	6.73	3.09
std	12.88	1.24	91.00	423.44	1.24
min	1.00	6.00	4.00	0.00	1.00
25%	26.00	8.00	98.00	0.01	2.00
50%	28.00	8.00	147.00	0.01	3.00
75%	39.00	8.00	218.00	0.02	4.00
max	119.00	36.00	605.00	2948.00	5.00

Pipe material considered included – VCP, PVC, RCP, CI, and DI. The full form of the pipe material is presented in Appendix Section A. The native soil type included in the dataset was found to be sand, sand fines, clay, and loam. Table 5-45 shows the unique values of the independent variable of the Northeast dataset. The dataset for developing the models was divided into 70%-train set and 30% test set.

Table 5-45 Unique values of Northeast Dataset

Pipe Age	78
Pipe Diameter	9
Pipe Material	8
Pipe Length	412
Pipe Slope	3047
Pipe Native Soil	4
PACP Score	5

5.5.1.2 Results on Developed Models

5.5.1.2.1 Decision Tree (DT) Classifier with Imbalanced Dataset

The confusion matrix developed for the DT classifier method with default parameters using the imbalanced data for the Northeast is illustrated in Figure 5-65. The ROC curve for the DT classifier is plotted and shown in Figure 5-66. The training set and test set performance is viewed in Table 5-46, and it is seen that the model displays poor Accuracy and F1- score. The confusion matrix shown below depicts that there are few misclassifications in classes 1 and 3. Also, although there are less misclassifications on classes 4 and 5, the true predicted values in those classes are lesser in number.

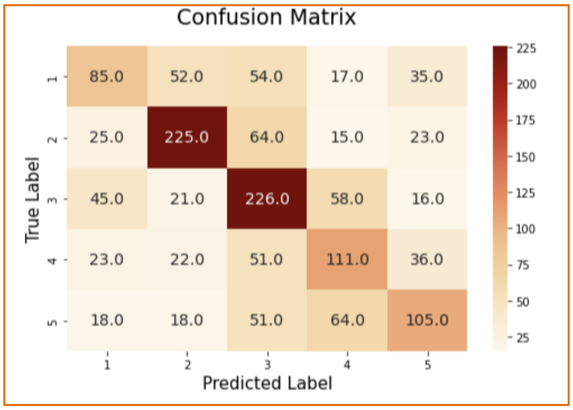


Figure 5-65 Confusion Matrix for Decision Tree Classifier–Northeast

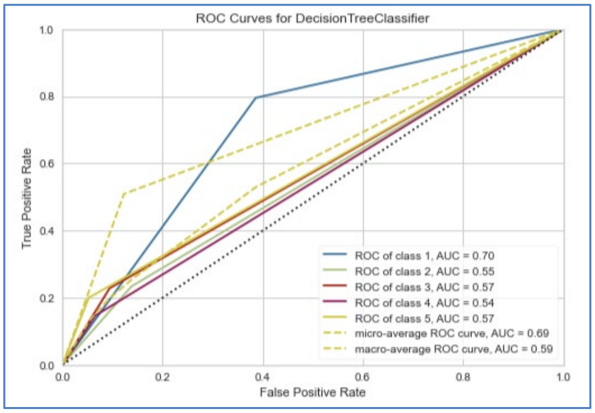


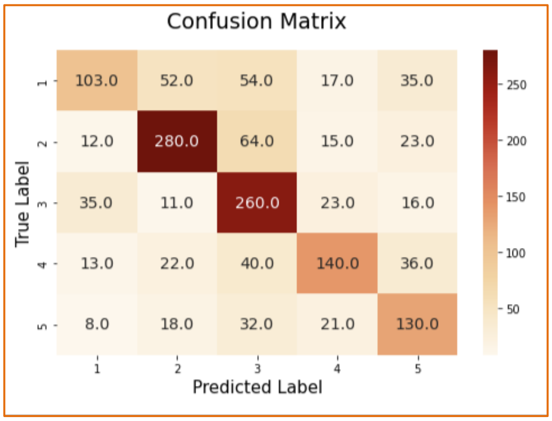
Figure 5-66 ROC Curve for Decision tree Classifier-Northeast

Table 5-46 Performance of Decision Tree Classifier–Northeast

Performance	Accuracy	Macro F1-score
Training set	100.00%	99.72%
Test set	52.77%	53.88%

5.1.2.2 Random Forest (RF) Classifier Method with Imbalanced Data

The confusion matrix for the RF classifier with default using imbalanced data for Northeast is shown in Figure 5-67. The ROC curve for this model is plotted in Figure 5-68. The performance of the training and test data can be reviewed in Table 5-47. This model is shown to be overfitting.



	1	2	3	4	5
1	103.0	52.0	54.0	17.0	35.0
2	12.0	280.0	64.0	15.0	23.0
3	35.0	11.0	260.0	23.0	16.0
4	13.0	22.0	40.0	140.0	36.0
5	8.0	18.0	32.0	21.0	130.0

Figure 5-67 Confusion Matrix for Random Forest Classifier–Northeast

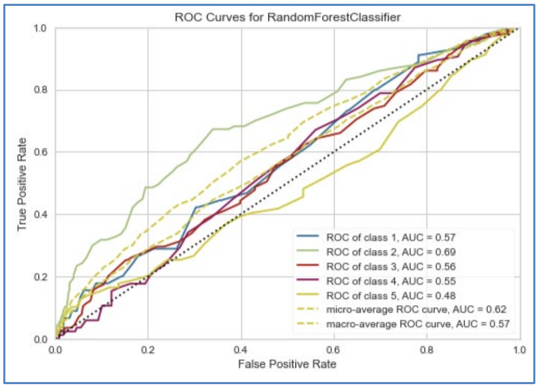


Figure 5-68 ROC Curve for Random Forest Classifier–Northeast

Table 5-47 Performance of Random Forest Classifier-Northeast

Performance	Accuracy	Macro F1-score
Training set	99.86%	99.86%
Test set	56.23%	56.02%

5.5.1.2.3 Tuned Decision Tree (DT) Classifier using Over-sampled Data

The grid parameters for the tuned DT Classifier are represented in Table 5-48. The confusion matrix for the tuned DT classifier is presented in Figure 5-69, and it depicts that there is a reduction in the misclassification in the values which are predicted non-diagonal area of the matrix. The ROC curve is plotted for the model in Figure 5-70. The performance values of this model are listed in Table 5-49.

Table 5-48 Tuned Decision Tree Classifier Grid Parameters

max_depth	35
max_leaf_nodes	15
min_samples_leaf	5
random_state	1

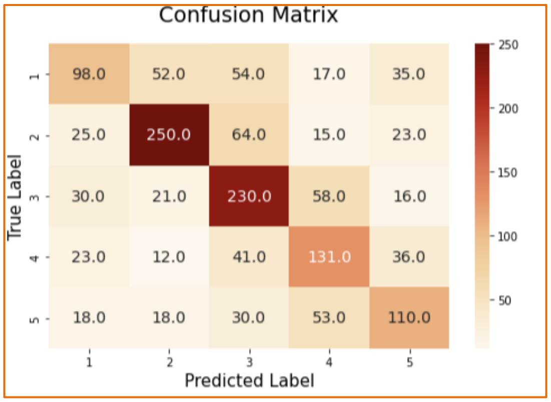


Figure 5-69 Confusion Matrix for Tuned Decision Tree Classifier-Northeast

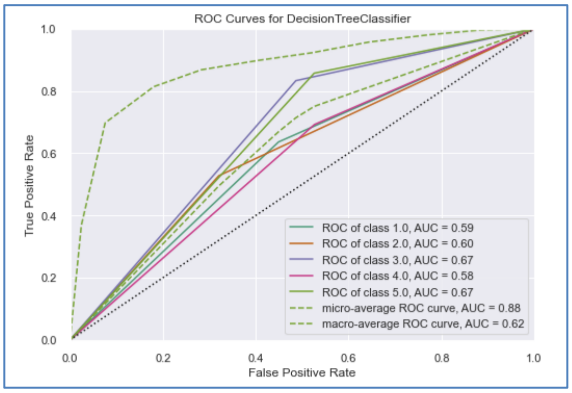


Figure 5-70 ROC Curve for Tuned Random Forest Classifier-Northeast

Table 5-49 Performance of Tuned Decision Tree Classifier-Northeast

Performance	Accuracy	Macro F1-score
Training set	58.54%	58.40%
Test set	58.04%	58.05%

5.5.1.2.4 Tuned Random Forest (RF) Classifier using Over-sampled Data

The confusion matrix for tuning of RF method is listed below and the grid parameters for this model are shown in Table 5-50. The tuned RF classifier confusion matrix is shown in Figure 5-71. The ROC curve for this model is plotted and shown in Figure 5-72. The training set and test set performance for accuracy and F1-score is viewed in Table 5-51. This model is not overfitting.

Table 5-50 Tuned Random Forest Grid Parameters

max_depth	10
max_samples	0.2
min_impurities_decrease	0.0001
N_estimates	150
random_state	1

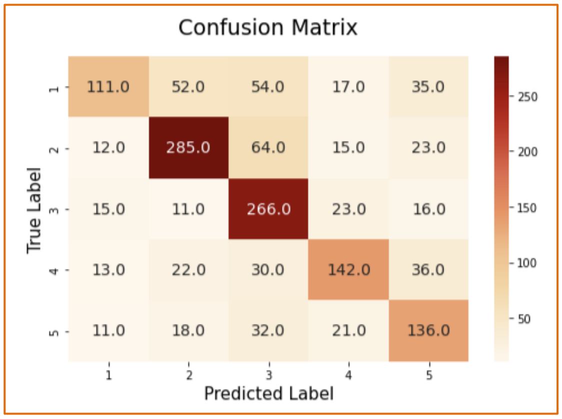


Figure 5-71 Confusion Matrix for Tuned Random Forest Classifier-Northeast

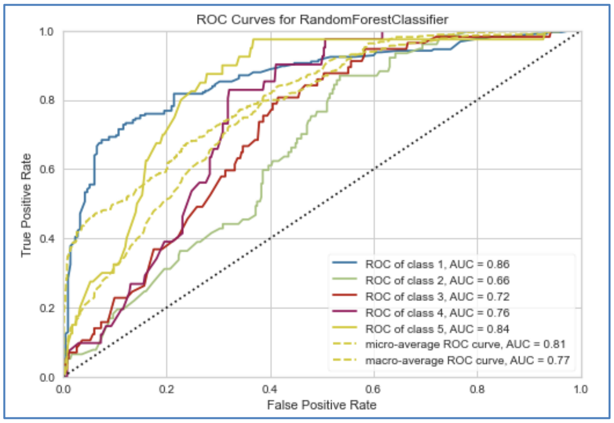


Figure 5-72 ROC Curve for Tuned Random Forest Classifier-Northeast

Table 5-51 Performance of Tuned Random Forest Classifier – Northeast

Performance	Accuracy	Macro F1-score
Training set	68.12%	70.35%
Test set	67.02%	67.17%

5.5.1.2.5 AdaBoost Classifier with Default Parameters using Imbalanced Data

The confusion matrix for the AdaBoost classifier with default parameter is illustrated in Figure 5-73. The ROC curve for the AdaBoost classifier is plotted and shown in Figure 5-74. The training set and test set performance for accuracy and F1- score are listed in Table 5-52. This model is not overfitting but has poor accuracy score.

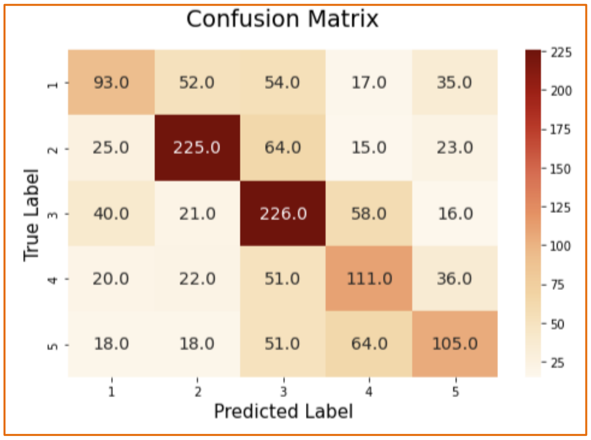


Figure 5-73 Confusion Matrix for AdaBoost Classifier-Northeast

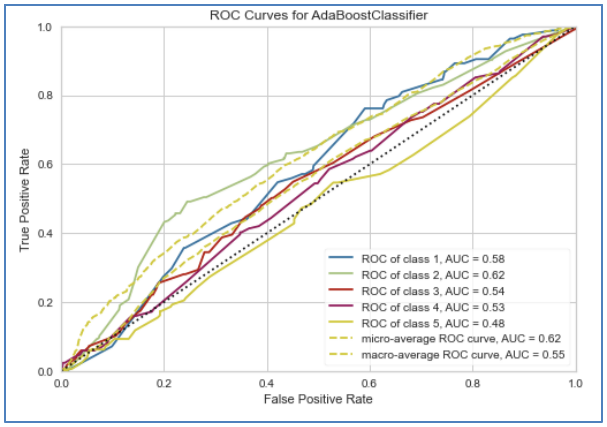


Figure 5-74 ROC Curve for AdaBoost Classifier-Northeast

Table 5-52 Performance of AdaBoost Classifier – Northeast

Performance	Accuracy	Macro F1-score
Training set	57.25%	57.44%
Test set	55.28%	55.99%

5.5.1.2.6 Gradient Boost Classifier with Default Parameters using Imbalanced Data

The gradient boosting classifier with default parameters is executed on the given dataset of the Northeast, and its confusion matrix is presented in Figure 5-75. The ROC curve for the gradient classifier is visualized, as shown in Figure 5-76. The training set and test set performance for accuracy and F1- score is viewed in Table 5-53. This model is not overfitting and has better scores than AdaBoost classifier.

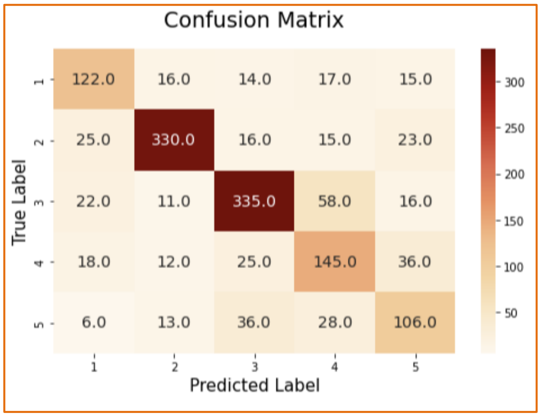


Figure 5-75 Confusion Matrix for Gradient Boost Classifier-Northeast

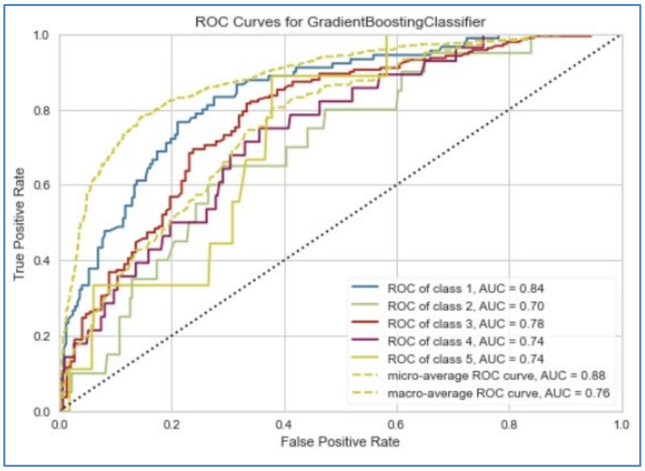


Figure 5-76 ROC curve for Gradient Boost Classifier-Northeast

Table 5-53 Performance of Gradient Boost Classifier-Northeast

Performance	Accuracy	Macro F1-score
Training set	68.66%	68.89%
Test set	67.99%	68.12%

5.5.1.2.7 XGBoost Classifier with Default Parameters using Imbalanced Data

The confusion matrix for the XGBoost classifier using default parameter is presented as shown in figure 5-77. The ROC curve for this model is shown in Figure 5-42. The training set and test set performance for accuracy and F1- score are listed in Table 5-54. This model is not overfitting and has better scores than other ensemble algorithms discussed above in this research.

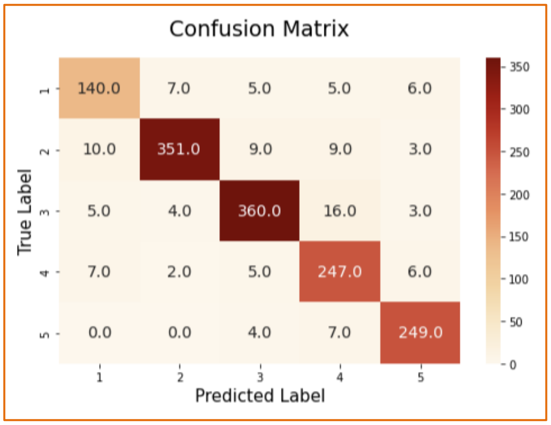


Figure 5-77 Confusion Matrix for XGBoost Classifier-Northeast

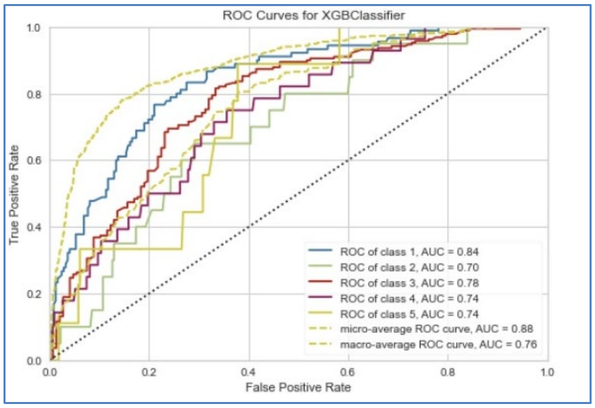


Figure 5-78 ROC Curve for XGBoost Classifier-Northeast

Table 5-54 Performance of XGBoost Classifier- Northeast

Performance	Accuracy	Macro F1-score
Training set	72.12%	72.88%
Test set	72.01%	72.65%

.5.1.2.8 Tuned AdaBoost Classifier using Over-sampled Data

The tuned Adaboost model was developed with over-sampled dataset to improvise the model suffering from imbalanced data. This model was tuned by using gridsearch CV and randomizedsearch CV. The confusion matrix for both the models is presented in Figure 5-77(a) and 5-77(b) below. It can be seen that a good number of values are predicted well on the diagonal element of matrix but, also there is misclassification of the values in all the classes which will affect the model performance. The ROC curve for this both gridsearch CV and randomizedsearch CV is shown in model is illustrated in Figure 5-78(a) and 5-78(b) and it is seen in Figure 5-

78(a) that area under curve for classes 2 and 5 is better than other classes and in Figure 5-78(b) the area under curve for classes 3 and 5 are better than other classes. The training set and test set performance for accuracy and F1- Score are listed in Table 5-55. This model is overfitting and has poor performance.

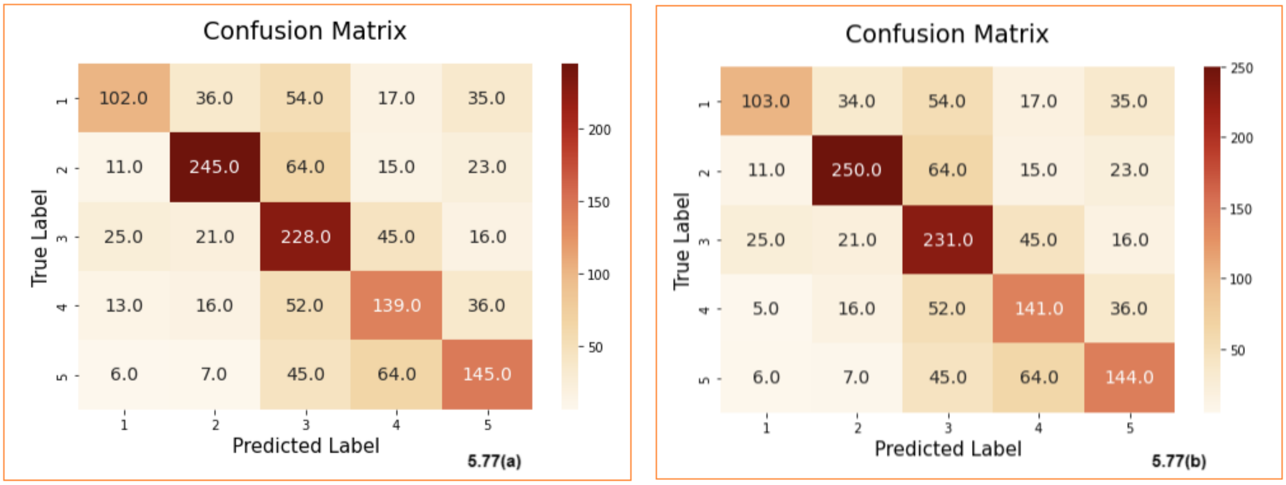


Figure 5-77(a) and 5-77(b) Confusion Matrix for Tuned AdaBoost Classifier with Gridsearch CV and Randomizedsearch CV-Northeast

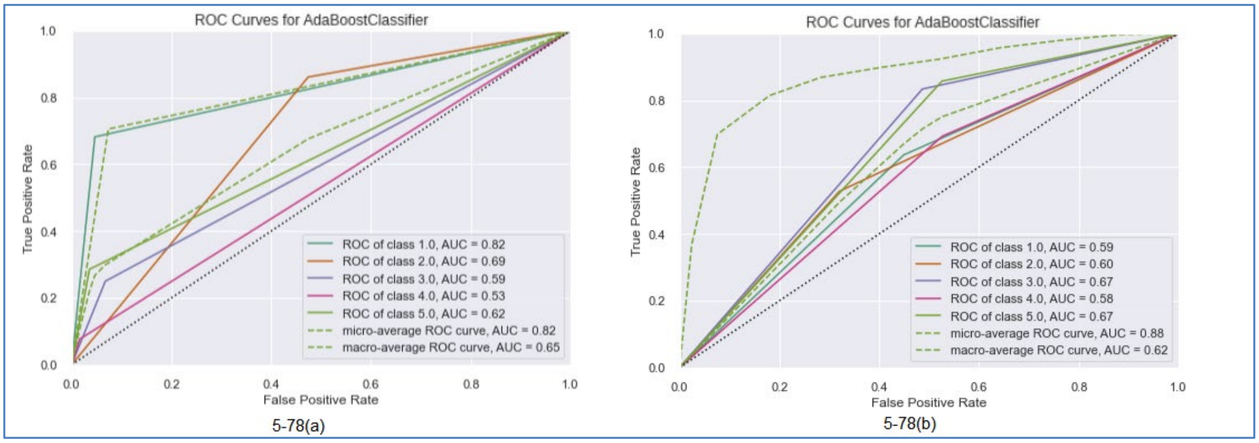


Figure 5-78(a) and 5-78(b) ROC Curve for Tuned AdaBoost Classifier with Gridsearch CV and Randomizedsearch CV-Midwest

Table 5-55 Performance of Tuned AdaBoost Classifier – Northeast

Performance	Accuracy	Macro F1-score
Gridsearch CV Training set	88.74%	88.74%
Gridsearch CV Test set	62.5%	62.85%
Randomizedsearch CV Training set	92.80%	92.62%
Randomizedsearch CV Test set	67.55%	67.25%

5.5.1.2.9 Tuned Gradient Boost Classifier using Over-sampled Data

The gradient boosting algorithm was tuned with hyperparameters with gridsearch CV and randomizedsearch CV, and the confusion matrix for those models is illustrated in Figures 5.79(a) and 5.79(b). The predicted value on the diagonal area is evidence that the model performance is rightly evaluated. For the 4 and 5 classes wrongly, predicted values are accounting to 0 which is a good, developed model. It is evident that the confusion matrix developed for these models is better than AdaBoost models as there is less misclassification in all classes and the higher values predicted rightly on diagonal element. The ROC curve for those models is plotted in Figures 5-80(a) and 5-80(b). The ROC curve shown in 5-80(a) has high TPR for the classes 1,2,3 and low TPR for classes 4 and 5 , this proves that the values shown in confusion for this model in Figure 5-79(a) has misclassification in classes 4 and 5.Similarly, the ROC curve in Figure 5-80(b) shows an excellent values for AUC under all the classes and the hence the performance for gradient boost classifier with randomizedsearch CV has better than the other.

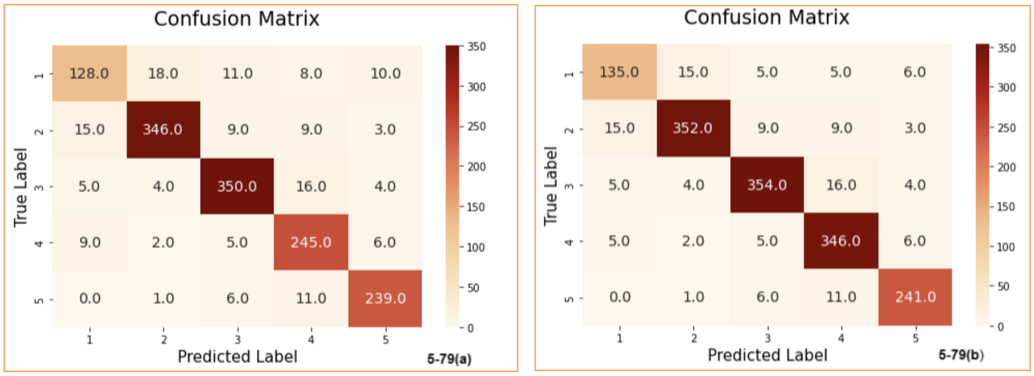


Figure 5-79(a) and 5-79(b) Confusion Matrix for Gradient Boost Classifier with Gridsearch CV and Randomizedsearch CV–Northeast

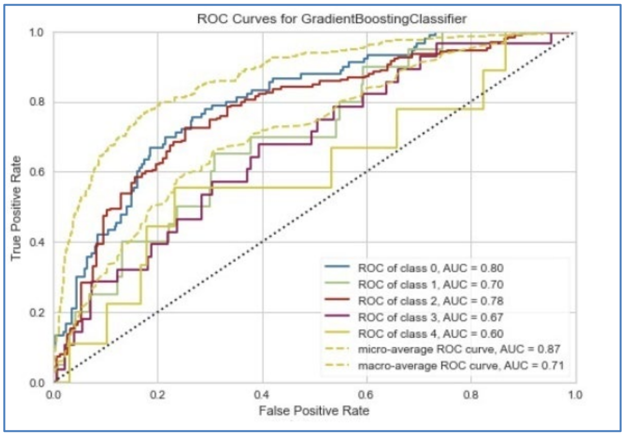


Figure 5.80(a) ROC Curve for Tuned Gradient Boost Classifier with Gridsearch CV – Northeast

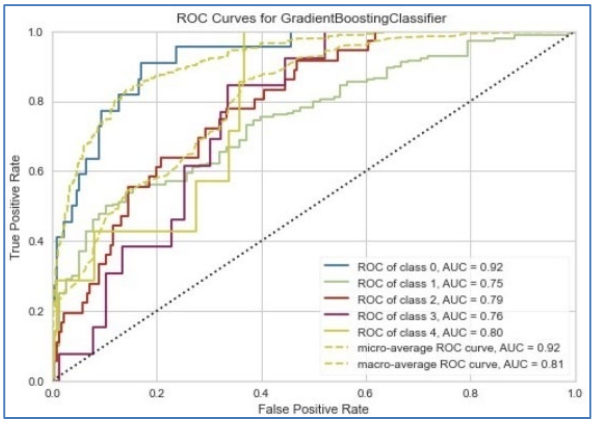


Figure 5.80(b) ROC Curve for Tuned Gradient Boost Classifier with Randomizedsearch CV – Northeast

Table 5-56 Performance of Tuned Gradient Boost Classifier-Northeast

Performance	Accuracy	Macro F1-score
Gridsearch CV Training set	74.25%	74.35%
Gridsearch CV Test set	72.85%	72.60%
Randomizedsearch CV Training set	78.12%	79.99%
Randomizedsearch CV Test set	77.33%	77.32%

5.5.1.2.10 Tuned XGBoost Classifier using Over-sampled Data

As the last model, XGBoost Classifier with hyperparameter tuning with respect to gridsearch CV and randomizedsearch CV was assessed in this study under the Northeast. The XGBoost method works similarly to the gradient boost classifier method but with much faster execution. The confusion matrix for both optimized search models and the ROC curves were developed and illustrated in Figures 5-81(a), 5-81(b), 5-82(a), and 5-

82(b). The training set and test set performance for accuracy and F1- score are tabulated under Table 5-57. This model is not overfitting and has better scores than all the algorithms discussed above in this research.

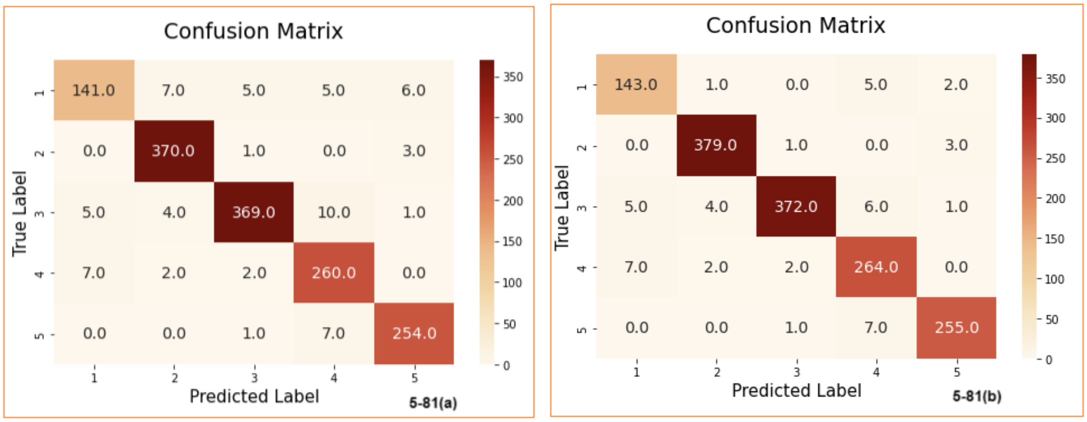


Figure 5-81(a) and 5-81(b) Confusion Matrix for XGBoost Classifier with Gridsearch CV and Randomizedsearch CV–Northeast

In the confusion matrix shown above, it shows that there are few misclassifications under all the classes and the diagonal elements as has much higher values predicted rightly compared to the confusion matrix in Figure 5-81(a) which indicates that model performance of randomizedsearch CV will better than the gridsearch CV.

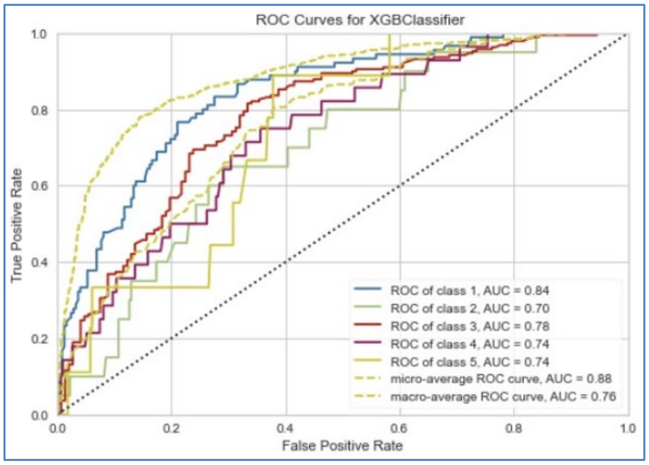


Figure 5-82(a) ROC Curve for Tuned XGBoost Classifier with Gridsearch CV- Northeast

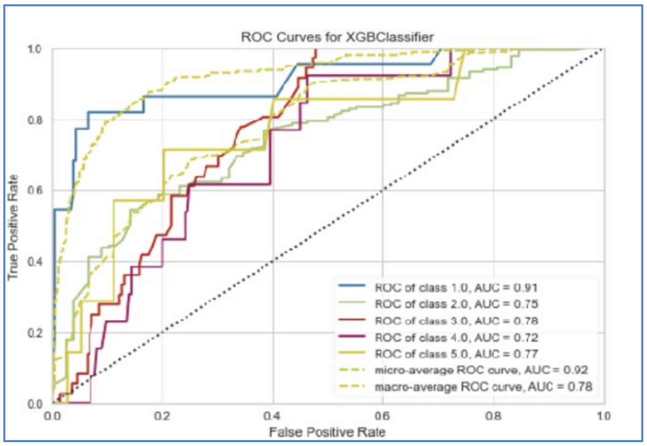


Figure 5-82(b) ROC Curve for Tuned XGBoost Classifier with Randomizedsearch CV- Northeast

The AUC for gridsearch CV has an overall good score for all classes above 70% with macro average of 76%, which is good indication that the values are predicted rightly. Similarly, AUC for randomizedsearch CV has a macro average of 78% which shows that randomizedsearch CV is the better model than gridsearch CV.

Table 5-57 Performance of XGBoost Classifier-Northeast

Performance	Accuracy	Macro F1-score
Gridsearch CV Training set	80.23%	80.35%
Gridsearch CV Test set	76.65%	76.52%
Randomizedsearch CV Training set	81.12%	79.98%
Randomizedsearch CV Test set	80.98%	79.08%

5.5.1.3 Discussion on the Developed Models – Northeast

Figures 5-83 and 5.84 summarize the feature importance and test performance of all the models developed for the Northeast. It can be observed that the models developed with over-sampled dataset with hyperparameters tuning yielded with good performance above 70%. Further, the gradient boost classifier with randomizedsearch CV (77%) and XGBoost classifier with gridsearch CV (77%) and randomizedsearch CV (79%) performed the best with randomizedsearch CV method.

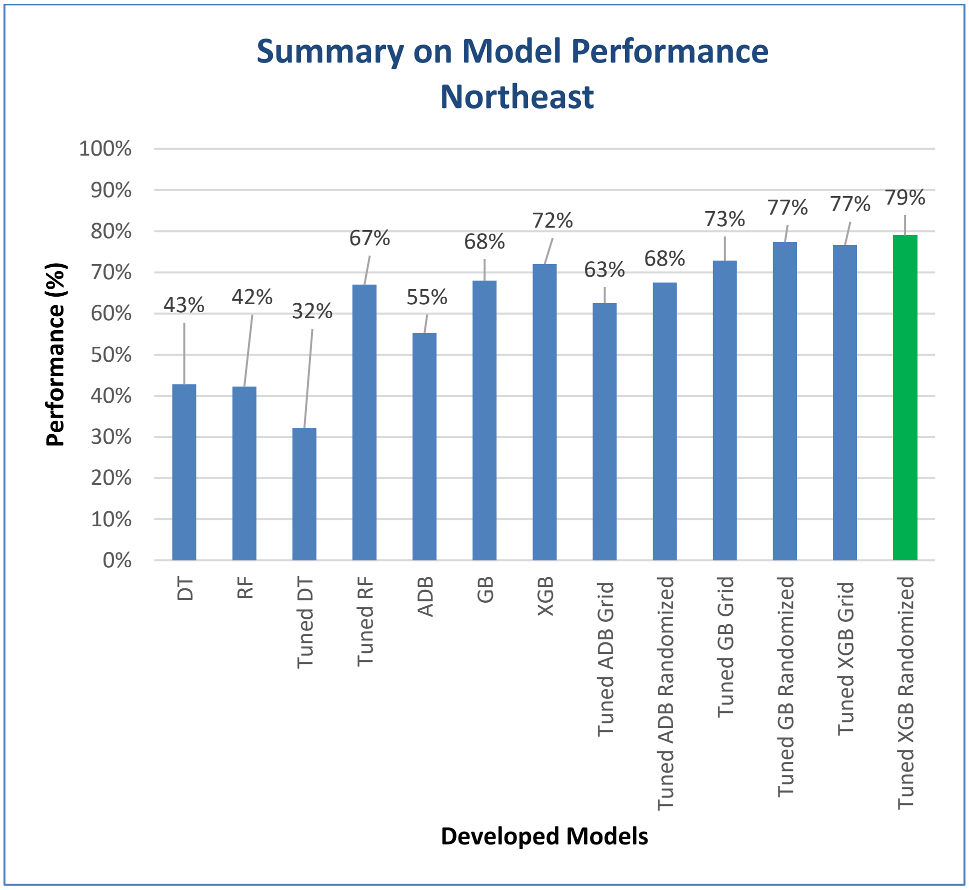


Figure 5-83 Summary on Models Test Performance-Northeast

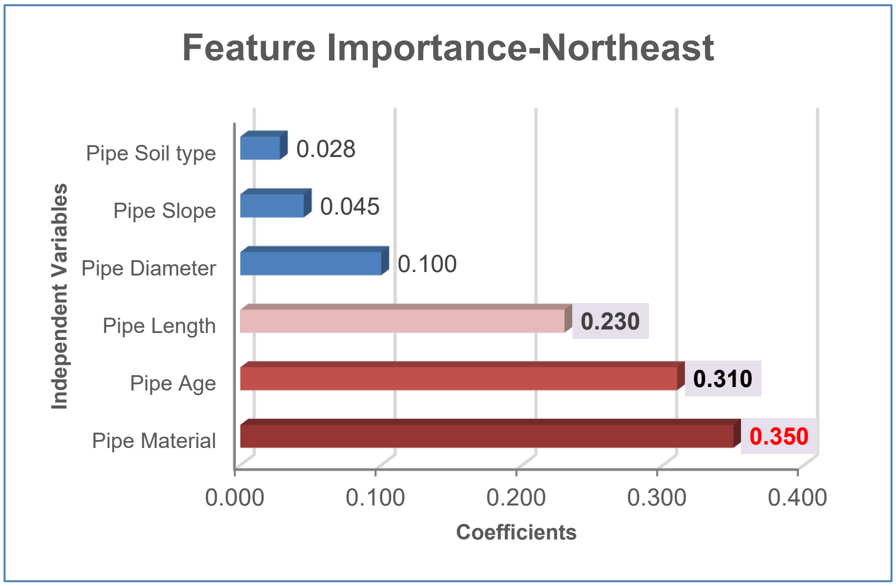


Figure 5-84 Summary on Feature Importance-Northeast

As the test performance of XGBoost based on Randomizedsearch CV algorithm is performed better than all the other models, the feature importance of the variables for this model was developed and is presented Figure 5-22. From this feature important observation, we can conclude that pipe material is the high impact variable with 0.350 coefficient value for the PACP score. Pipe age, pipe length, pipe diameter are the next most impacting variables on PACP score with coefficients of 0.310, 0.230 and 0.100, respectively. Pipe slope and pipe native soil type has the least coefficient values with 0.045 and 0.028 and can be concluded that they have the least effect on the PACP score. In order to evaluate on this conclusion, XGBoost model was reconstructed to check its performance by including the pipe age, pipe material, pipe diameter and pipe length variables in the dataset and it was found that, the macro F1-score for this model was 0.735 and this performance is close to the previous model performance constructed using all the independent variables. Therefore, it can be decided that the critical variables mentioned above in this section can be considered to have more effect on the PACP score.

5.3 Chapter Summary

Chapter 5 discusses the results developed on all the machine learning tree-based models by constructing confusion matrix, ROC curves and tabulating their accuracy and macro F1-score to compare the performance of each machine learning model. Based on the confusion matrix developed for all models, it was found that the models trained with imbalanced dataset failed to classify structurally poor condition pipes and the ones modelled with over-sampled data performed well. XGBoost tuned with gridsearch CV and randomizedsearch CV and gradient boost tuned with randomizedsearch CV had better performances than all other models.

CHAPTER 6

CONCLUSIONS, LIMITATIONS, PRACTICAL APPLICATIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

6.1 Conclusions

The below inferences were drawn from the development of prediction models discussed in the Chapter 5 for US regions – Southeast, Southcentral, Northeast and Midwest and conclusion on each model developed in this research is discussed distinctly for improved understanding on their performances.

6.1.1 Data Processing and Model Development.

Sewer pipe inspection data was collected from eleven city municipalities classified under the four US regions as mentioned above to develop a comprehensive prediction model for each region. It was found that this research had the most distinctive dataset collected when compared to previous studies which had sewer pipe data collected for one city. Also, the dataset collected from all the regions showed the major imbalance under all the 5 classes of PACP score especially pipe segments having 4 and 5 score were low in number. This imbalance in dataset was treated by adopting the technique SMOTE. The models were developed using imbalance data and over-sampled data to evaluate the variations in their performance. The best prediction model is selected under each region based on their high-performance score and the feature importance for that selected model is developed to see which independent variable is having highest impact on the target variable i.e., is PACP score.

6.1.2 Conclusion Summary on the Developed Prediction Models

Table 6.1 Conclusion Summary

#	Prediction Models	Dataset/Parameter	Accuracy / F1 Range	Inference/ Comments
1	Decision Tree Classifier	Imbalance / Default	52%-58%	The Model is overfitting and did not perform well as there was lot of misclassifications between predicted values and true values for the given dataset under each region considered in this research.
2	Random Forest Classifier	Imbalance / Default	56%-67%	The model was trained well but the test performance was not satisfactory. Like DT classifier, it was found that their misclassification in predicted values. So, the model is concluded as overfitting in nature.

3	Tuned Decision Tree Classifier	Imbalance / Default	54%-60%	The Model is not overfitting that means the model was trained well by Tuned the algorithm but did not perform well for the sewer pipe dataset.
4	Tuned Random Forest Classifier	Imbalance / Default	67%-73%	The Model is not overfitting that means the model was train well by Tuned the algorithm and model performance showed better performance when compared to other logistic regression models.
5	AdaBoost Classifier	Imbalance / Default	53%-66%	The model is not overfitting. The training and test performance showed that the model has low accuracy score. The confusion matrix developed are the regions for this model showed that non diagonal values are low in number depicting that there was misclassification of data.
6	Gradient Boost Classifier	Imbalance / Default	64%-74%	The model is not overfitting. The training and test performance showed the model has a low accuracy score. The confusion matrix developed for this model under all US regions showed that non-diagonal values were better than the AdaBoost model. But still, the performance of this model is not satisfactory.
7	XG Boost Classifier	Imbalance / Default	72%-77%	The model is not overfitting. The training and test performance showed that the model has a good accuracy score. The confusion matrix developed for this model under all US regions showed that diagonal and non-diagonal values were found better than all the models develop with default parameter.
8	Tuned AdaBoost Classifier with Grid-Search CV	Oversampled / Hyper parameter	64%-67%	The model is overfitting. The training and test performance showed that the model has a good accuracy score compared to the AdaBoost model developed with default parameters. The confusion matrix developed for this model for all the US regions showed that diagonal and non-diagonal values were found to be better than the Adaboost model developed with default models. But the model tuned with randomized search CV showed better performance compared to Grid-Search CV.
9	Tuned AdaBoost Classifier with Randomized Search CV	Oversampled / Hyper parameter	63%- 68%	
10	Tuned Gradient Boost Classifier with Grid-Search CV	Oversampled / Hyper parameter	72%-76%	The model is not overfitting. The training and test performance showed that the model has a good accuracy score compared to the AdaBoost model developed with default parameters and the gradient Boost model developed with default parameters. The confusion matrix developed for this model for all the US regions showed that diagonal and non-diagonal values were found to be better than the Adaboost model developed with default parameters, hyperparameter Tuned and gradient boost with default parameters. Also, the model developed with randomized search CV showed better performance than Grid-Search CV.
11	Tuned Gradient Boost Classifier with Randomized Search CV	Oversampled / Hyper parameter	76%-77%	
12	Tuned XG Boost Classifier with Grid-Search CV	Oversampled / Hyper parameter	73%-77%	The model is not overfitting. The training and test performance showed that the model has the best accuracy score compared to all the models developed in this study. The confusion matrix

13	Tuned XG Boost Classifier with Randomized Search CV	Oversampled / Hyper parameter	76%-79%	developed for this model for all the US regions showed that diagonal and non-diagonal values were found to be better and predicted the most accurate true positive values. Also, the model developed with randomized search CV illustrated that it is the best approach to develop the prediction model.

6.1.3 Best Model Selected

As per the discussion on results presented in Chapter 5 and from the conclusion summary illustrated in Table 6.1, it can be inferred that XG Boost classifier with over-sampled hyperparameter tuning using randomizedsearch CV is the best approach to develop the prediction model for the sewer pipe inspection data collected for this research. Also, the second-best model which has performed better is found to be gradient boost classifier with over-sampled hyperparameter with randomizedsearch CV.

6.1.4 Feature Importance

The feature importance was plotted for the best approach, and it was shown in the Chapter 5 results sections that the independent variable in the order of pipe age, pipe material, pipe diameter and pipe length had the better TPR rate and had the highest impact on the PACP score and the variables like soil slope and native soil type had low TPR and showed the least impact on the PACP score for this research. This means that, if the variables like native soil and slope of the pipe are removed from the dataset, and the model is built by including the independent variables having high impact like pipe age, pipe material, pipe diameter and pipe length, the performance of the model will not be affected much.

6.1.5 Comparison and Conclusion on the Critical Variables Identified Under Each US Region

US Region	**Best Model Selected**	**Critical Factors Identified**	**Feature Importance Coefficients**	**Weighted F1 score Model 1**	**Weighted F1 score Model 2**	**Comments**
Southeast	Tuned XGBoost with Over-sampled using Randomized SearchCV	Pipe Age	0.445	0.786	0.756	The weighted F1-score of models 1 and 2 shows that the performance of both models is almost similar. This concludes that the critical variables shown in this table for Southeast are rightly identified and the city municipalities can use the developed model as an approach for future inspection of the pipes for the area under study. It was observed from the sewer pipe dataset that pipes under PACP score of 4 and 5 for pipe age, material and diameter was found to be as 45 years to 71 years., PVC and UnReinCONC and 8 – 15 inches, respectively.
		Pipe Material	0.25			
		Pipe Diameter	0.21			
Southcentral	Tuned XGBoost with Over-	Pipe Age	0.36	0.788	0.800	The weighted F1-score of models 2 is 0.80 which is more than that of model 1 which is 0. 788.This better performance of model 2 concludes that, performance

	sampled using Randomized SearchCV	Pipe Material	0.37			of both the model almost matched and the identified critical variables shown in this table for Southcentral is rightly identified and the city municipalities can use the developed model as an approach for future inspection of the pipes for the area under study. It was observed from the sewer pipe dataset that pipes under PACP score of 4 and 5 for pipe age, material and diameter was found to be as 45 years to 85 years., UnReinCONC and VCP and 24 – 54 inches, respectively.
		Pipe Diameter	0.2			
Midwest	Tuned XGBoost with Over-sampled using Randomized SearchCV	Pipe Age	0.15	0.769	0.69	The weighted F1-score of models 1 and 2 shows that the performance of both models is not similar. This shows that the sewer pipe data used in the model building is not sufficient to optimize the model's performance and more inspection data is needed to collect in future to identify the critical variables. It was observed from the sewer pipe dataset that pipes under PACP score of 4 and 5 for pipe age, material and diameter was found to be as 51 years to 72 years., RPM and VCP and 6 – 54 inches, respectively.
		Pipe Diameter	0.11			
		Pipe Material	0.09			
Northeast	Tuned XGBoost with Over-sampled using Randomized SearchCV	Pipe Material	0.35	0.786	0.74	The weighted F1-score of models 1 and model 2 shows that the performance of both models are almost similar. This concludes that the critical variables shown in this table for Southeast are rightly identified and the city municipalities can use the developed model as an approach for future inspection of the pipes for the area under study. It was observed from the sewer pipe dataset that pipes under PACP score of 4 and 5 for pipe age, material and diameter was found to be as 25 years to 72 years., PVC and UnReinCONC and 8 – 12 inches, respectively.
		Pipe Age	0.31			
		Pipe Length	0.23			
		Pipe Diameter	0.1			
Model 1-Model developed includes all the independent variables						
Model 2-Model developed includes only the identified critical independent variables						

6.2 Limitations

As indicated previously, this research is undertaken to demonstrate the prediction model developed using tree-based machine learning algorithms which are expected to produce better model performance. The main limitation of condition prediction models is the availability of appropriate datasets to generate the models. Environmental parameters affecting the condition of sanitary sewer pipes, such as bedding material, overburden pressure, soil water content, traffic flow, and other factors identified in the literature, were omitted due to a lack of a proper dataset. On the other hand, the population of sanitary sewer pipes in condition levels 1,2 and 3 was found to be more in number in the given dataset for the region with respect to levels 4 and 5. This variable in the dataset caused low performance in developed models like random forest, which has always proved to be one of the efficient approaches to build a prediction model.

6.3 Contribution to the Body of Knowledge

The major contributions of this research are listed below.

- Several machine learning models were constructed in previous research to predict the condition of the sewer pipes. This research added high performing ML algorithms which are known to produce the efficient prediction model. Techniques like hyperparameter tuning were also included in this study to improve the performance of the models.
- In this study, the sewer pipe data was collected from eleven cities all over US regions, this is one of the outstanding achievements of this study where no other single research has collected the data from more than two cities to develop the prediction model to access the future condition of sewer pipe.

6.4 Practical Applications

Sewer pipe inspections are crucial for maintaining the functionality and integrity of the sewer system. To ensure consistent and effective inspections, municipalities should establish clear rules and guidelines. According to Environmental Protection Agency (EPA) as of September 2021 has established the Clean Water Act (CWA) to rule out the framework for regulating discharges of pollutants into U.S. waters and sets water quality standards, besides there was no specific nationwide order issued by the United States regarding sewer pipe inspection. Under the CWA, the EPA works with states to implement and enforce various programs, including the National Pollutant Discharge Elimination System (NPDES) and the Municipal Separate Storm Sewer System (MS4) program. Most of the city municipalities set up the inspection of sewer pipes program annually to keep in check on the proper functioning of the sewer pipes.

Sanitary sewage utility managers may have to consider prediction models while making decisions about the rehabilitation and replacement of sewer lines. Since assessing the state of sanitary sewers and gathering data through CCTV inspection are time-consuming and expensive. It is impossible to inspect every sewer pipe within a city's boundaries annually. Additionally, only around one-third of sanitary sewage systems are examined every five years due to accessibility issues and a lack of financing.

Prediction models may be influential to sanitary sewer utilities managers in the decision-making process in rehabilitation and replacement of sewer pipes. Since sanitary sewer condition assessment and data collection through CCTV inspection are time-consuming and are proven too expensive. Inspecting all the pipes under a city limit is one impossible task. Also, Due to inaccessibility and inadequate funding, only about one third of sanitary sewer systems are inspected every 5 years.

Prediction models can assist in expediting the evaluation of the condition rating of sewer pipes using independent variables. This research can be able to assist the city managers and utility owners to develop a framework to construct a prediction model methodology that can be employed to their sewer pipe system prevailing in their area. The dissertation gives out a complete procedure on the data processing and the model development which can be followed to create the specific framework that works for that utility.

The final selected best prediction model in this research can be used as the base support tool in the future case studies to evaluate the sewer pipe conditions. For example, considering the prediction modelling for Southcentral, it was found that the XGBoost hyperparameter tuned model was selected as the best algorithm as it showed the good accuracy score of 79%. And the feature importance for this model was found to be pipe in the order of pipe 1) material (0.37), 2) age (0.36) ,3) diameter (0.20) 4) length (0.075) 5) slope (0.065) and native soil (0.031) having effect on the PACP score. Once the critical variables are identified through this process, the pipe segments with top 4 critical variables (excluding the variables with low feature importance) are used to develop the prediction model to validate its performance with the previous produced model. For this case study, the overall weighted F1-score for the model including all independent variables (model 1) and the model including only first three critical variables were found to be 0.79 and 0.80, respectively. This implies that the performances of both models are similar, and the critical variables are rightly identified. So, for the south-central region, the city managers can target the critical variables under the PACP score of 4 and 5 for the future inspection process instead of performing the mass inspection on all the pipes. This dissertation could be used in the decision-making process for replacement or rehabilitation and inspection prioritization too. Upon request, the research models and code can be shared with interested city managers and utility owners.

6.5 Recommendations for Future Research

This dissertation can be expanded by considering the below observations.

- Dataset with other independent variables, such as backfill type, bedding material, soil moisture, overburden pressure, installation method, pipe shape, previous maintenance, overflow, and blockage history, can be focused on to be collected to improve the deterioration models developed in this dissertation.

- Also, this study was not able to collect the sewer pipe dataset from western US regions. So, including the sewer pipe data from the west US region can have a more comprehensive approach in comparing the prediction models.
- Deep learning algorithms can be investigated to develop condition prediction models and may act crucial to future research.
- Cost-benefit analysis could be accomplished to examine the cost savings for the municipality or utility owners.
- Sewer inspection data collected using the portable robots with cameras needs to be considered for future research as this is one of the latest emerging technologies for assessing the sewer pipe conditions with better accuracy.

CHAPTER 7

REFERENCES

American Water Works Association (AWWA), (2012). Buried no longer: confronting America's water infrastructure challenge, Boulder, CO, Retrieved from: https://www.awwa.org/Portals/0/files/legreg/documents/BuriedNoLonger.pdf, (2012)

Ariaratnam, S. T., El-Assaly, A., and Yang, Y. (2001). “Assessment of Infrastructure Inspection Needs Using Logistic Models.” Journal of Infrastructure Systems, 7(4), 160–165.

ASCE (2021). 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.

Bakry, I., Alzraiee, H., Kaddoura, K., Masry, M. E., and Zayed, T. (2016). “Condition Prediction for Chemical Grouting Rehabilitation of Sewer Networks.” Journal of Performance of Constructed Facilities, 30(6), 04016042.

Baur, R., and Herz, R. (2002). “Selective inspection planning with ageing forecast for sewer types.” Water Science and Technology, 46(6-7), 389–396.

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., Pitt, R., Nix, S. J., and Durrans, S. R. (2000). “Urban Wastewater Management in the United States: Past, Present, and Future.” Journal of Urban Technology, 33–62.

Chae, M. J. and Abraham, D. M. (2001). “Neuro-fuzzy approaches for sanitary sewer pipeline condition assessment.” Journal of Computing in Civil Engineering, 15(1), 4 – 14.

Charniak, E. and McDermott, D. (1985). Introduction to Artificial Intelligence. Addison- Wesley.

Chughtai, F. R. M. (2008). “Integrated condition assessment models for sustainable sewer pipelines.” thesis.

Chughtai, F., and Zayed, T. (2001). “Integrating WRc and CERIU Condition Assessment Models and Classification Protocols for Sewer Pipelines.” Journal of Infrastructure Systems, 17(3), 129–136

Coles, S. (2011). An introduction to statistical modeling of extreme values. Springer, London.

Dasu, T., and Johnson, T. (2003). Exploratory data mining and data cleaning. John Wiley & Son, Hoboken, NJ

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 73-89.

EPA (2002). The Clean Water and Drinking Water Infrastructure Gap Analysis, EPA-816- R-02-020. Washington DC: US Environmental Protection Agency, Office of Water.

EPA, (2004), “Report to Congress; Impacts and Control of CSOs and SSOs.” August 2004, <www. Ep.gov/nrmrl>.

EPA, (2007), “Innovation and Research for Water Infrastructure for the 21st Century.” April 2007, <www. Ep.gov/nrmrl>.

EPA, (2009), “Condition Assessment of Wastewater Collection Systems.” March 2009, <www. Ep.gov/nrmrl>.

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

Fenner, R. A. (2000). Approaches to sewer maintenance: a review. Urban Water, 343-356.

Freund Y., Schapire R., 1999. A Short Introduction to Boosting. Journal of Japanese Society for Artificial Intelligence, 14(5), 771-780

Geron, A. (2017). Hands-On Machine Learning with Scikit-Learn and TensorFlow. Sebastopol: O'Reilly.

Gheytaspour, M., & Habibzadeh Bigdarvish, O., (2018). Forecasting Oxygen Demand in Treatment Plant Using Artificial Neural Networks. International Journal of Advanced Engineering Research and Science, 5(3), 050-057. http://dx.doi.org/10.22161/ijaers.5.3.8

Hahn, M., Palmer, R., Merrill, S., and Lukas, A. (2002). Expert System for Prioritizing the Inspection of Sewers: Knowledge Base Formulation and Evaluation. Journal of Water Resources Planning and Management, 121- 129.

Harvey, R. R., and Mcbean, E. A. (2014a). “Predicting the structural condition of individual sanitary sewer pipes with random forests.” Canadian Journal of Civil Engineering, 41(4), 294–303.

Harvey, R. R., and Mcbean, E. A. (2014b). “Comparing the utility of decision trees and support vector machines when planning inspections of linear sewer infrastructure.” Journal of Hydroinformatics, 16(6), 1265–1279.

Hastie, T., Tibshirani, R., Friedman, J., (2017) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Hawari, A., Alkadour, F., Elmasry, M. and Zayed, T., (2016), “Simulation-Based Condition Assessment Model for Sewer Pipelines”, Journal of Performance of Constructed Facilities, 04016066.

Hernandez, N., Caradot, N., Sonnenberg, H., Rouault, P., and Torres, A. (2017). Support Tools to Predict the Critical Structural Condition of Uninspected Sewer Pipes in Bogota D.C.

The Leading-Edge Sustainable Asset Management of Water and Wastewater Infrastructure Conference. Trondheim, Norway.

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

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., and Abraham, D. M. (2005). "An ordered probit model approach for developing Markov chain-based deterioration model for wastewater infrastructure systems." Proc. Pipelines 2005: Optimizing Pipeline Design, Operations, and Maintenance in Today's Economy, Reston, VA., 649-661.

Kabir, G., Balek, N. B. C., and 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., and Guleria, S. P. (2015). "Geotechnical Investigation of Black Cotton Soils." International Journal of Advances in Engineering Sciences, Vol. 5, Issue 2, pp. 15- 22.

Khan, Z., Zayed, T., and Moselhi, O. (2010). "Structural Condition Assessment of Sewer Pipelines." Journal of Performance of Constructed Facilities, 24(2), 170–179.

Kley, G., and Caradot, N. (2013). "D1.2. Review of Sewer Deterioration Models." rep. Kompetenzzentrum Wasser Berlin gGmbH, Berlin, Germany, rep.

Koehn, P. (1994). Combining genetic algorithms and neural networks: the encoding problem.

Kulandaivel, G. (2004). "Sewer pipeline condition prediction using neural network models." thesis.

Kumar, S. S., Abraham, D. M., Jahanshahi, M. R., Iseley, T., and Starr, J. (2018). "Automated defect classification in sewer closed circuit television inspections using deep convolutional neural networks." Automation in Construction, 91, 273–283.

Laakso, T., Kokkonen, T., Mellin, I., and Vahala, R. (2018). "Sewer Condition Prediction and Analysis of Explanatory Factors." Water, 10(9), 1239.

Liashchynskyi, P. and Liashchynskyi, P. (2012). Gridsearch, Random-Search, Genetic Algorithm: A Big Comparison for NAS. arXiv:1912.06059.

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).

Lubini, A. T., and Fuamba, M. (2011). "Modeling of the deterioration timeline of sewer systems." NRC Research Press.

Luger, G. F. (2009). Artificial Intelligence: structures and Strategies for Complex Problem Solving. Addison Wesley, Harlow, England.

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., Tabesh, A., Riley, J. and Gruber, J. (2019). “Condition Prediction of Sanitary Sewer Pipes,” ASCE Pipeline Conference, 2019, Nashville, TN, U.S.

Mashford, J., Marlow, D., Tran, D., and May, R. (2011). “Prediction of Sewer Condition Grade Using Support Vector Machines.” Journal of Computing in Civil Engineering, 25(4), 283–290.

McDonald, S. E., Zhao, J. Q. (2001). “Condition Assessment and Rehabilitation of Large Sewers.” National Research Council Canada Report No. NRCC-44696. Ottawa, ON, Canada.

Micevski, T., Kuczera, G., and Coombes, P. (2002). “Markov Model for Storm Water Pipe Deterioration.” Journal of Infrastructure Systems, 8(2), 49–56.

Misiunas, D. (2005). Failure monitoring and asset condition assessment in water supply systems. Sweden: Lund University.

Najafi, M., and Gokhale, S. B. (2022). Trenchless technology: pipeline and utility design, construction, and renewal. McGraw-Hill, New York.

Najafi, M., and Kulandaivel, G. (2005). “Pipeline Condition Prediction Using Neural Network Models.” Pipelines 2005.

Najafi, M. (2016). Pipeline infrastructure renewal and asset management. McGraw-Hill Education, New York.

National Association of Sewer Service Companies (NASSCO), Pipeline Assessment Certification Manual, Dallas (2015).

Opila, M. C. (2011). “Structural Condition Scoring of Buried Sewer Pipes for Risk-Based Decision Making.” Doctoral Dissertation. University of Delaware. Newark, DE, USA.

Patel, H and Prajapati, P. (2018). “Study and Analysis of Decision Tree Based Classification Algorithms.” International Journal of Computer Sciences and Engineering. 2347-2693

Pyle, D. (2007). Data preparation for data mining. Morgan Kaufmann, San Francisco. Rahman, S. and Vanier, D.J (2004). NRC Report: An Evaluation of Condition Assessment Protocols for Sewer Management, National Research Council Canada (https://nparc.nrc-cnrc.gc.ca/eng/view/object/?id=93ed3e91-be5e-452d-b79d-c20d4ac77002).

Ross Quinlan, J., Bagging, boosting, and C4.5. In Proceedings, Fourteenth National Conference on Artificial Intelligence, 1996.

Salman, B. (2010). “Infrastructure management and deterioration risk assessment of wastewater collection systems.” Doctoral Dissertation. University of Cincinnati. OH, USA.

Salman, B., and Salem, O. (2012). “Modeling Failure of Wastewater Collection Lines Using Various Section-Level Regression Models.” Journal of Infrastructure Systems, 18(2), 146–154.

Seo, S. (2006). “A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets.” Thesis, Pittsburgh, PA, USA.

Singh, A., and Adachi, S. (2013). “Bathtub curves and pipe prioritization based on failure rate.” Built Environment Project and Asset Management, 3(1), 105–122.

Sousa, V., Matos, J. P., and 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.” dissertation.

Syachrani, S., Jeong, H. S. “D., and 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., and Dzuray, E. J. (2004). Sewer Pipeline Performance Indicators: Learning from the European Experience. Building Partnerships, Water resources 2000, 1-10.

Thornhill, R & Wildbore, P, (2005), “Sewer Defect Codes: Origin and Destination”, Utech Tran, D. H., Ng, A. W. M., Perera, B. J. C., Burn, S., and Davis, P. (2006). “Application of probabilistic neural networks in modelling structural deterioration of stormwater pipes.” Urban Water Journal, 3(3), 175–184.

Tran, D. H., Perera, B. J. C., and Ng, A. (2009). “Comparison of Structural Deterioration Models for Stormwater Drainage Pipes.” Computer-Aided Civil and Infrastructure Engineering, 24(2), 145–156.

Tran, D., Ng, A., and Perera, B. (2006). “Neural networks deterioration models for serviceability condition of buried stormwater pipes.” Engineering Applications of Artificial Intelligence, 20(8), 1144–1151.

Tran, H. D. (2007). “Investigation of Deterioration Models for Stormwater Pipe Systems.” Doctoral Dissertation. Victoria University. Melbourne, Australia.

Tran, D., Ng, A. W., Perera, B. J., Burn, S., and Davis, P. (2007). Application of probabilistic neural networks in modelling structural deterioration of stormwater pipes. Urban Water Journal, 175-184.

Tscheikner-Gratl, F., Caradot, N., Cherqui, F., Leitão, J. P., Ahmadi, M., Langeveld, J. G., . . . Clemens, F. (2020). Sewer asset management – state of the art and research needs. Urban Water Journal, 662-675.

Wasim, M., Shoaib, S., Mubarak, N. M., Inamuddin, and Asiri, A. M. (2018). “Factors influencing corrosion of metal pipes in soils.” Environmental Chemistry Letters, 16(3), 861–879.

Wirahadikusumah, R., Abraham, D., and Iseley, T. (2001). "Challenging issues in modeling deterioration of combined sewers." J. Infrastruct. Syst., 10.1061/(ASCE)1076-0342(2001)7:2(77), 77–84.

WRc, 1983 & 1986, Sewerage Rehabilitation Manual, Water Research Center, UK, London.

Wright, L. T., Heaney, J. P., and Dent, S. (2006). Prioritizing Sanitary Sewers for Rehabilitation Using Least-Cost Classifiers. ASCE JOURNAL OF INFRASTRUCTURE SYSTEMS, 174-183.

APPENDIX

Section A

Abbreviation

AC-Asbestos Cement

AI-Artificial Intelligence

ANN- Artificial Neural Network

ASCE- American Society of Civil Engineering

AUC- Area under the Curve

CCTV–Closed-circuit Television

CI-Cast Iron

CLC-Clay-lined Concrete Pipe

CMP-Corrugated Metal Pipe

CSS-Combined Sewer System

CV-Cross-Validation

CWA-Clean Water Act

DEC-Department of Environmental Conservation

DI- Ductile Iron

Downstream MH-Downstream Manhole

DT- Decision Trees

EDA-Exploratory Data Analysis

EPA- Environmental Protection Agency

FN- False Negative

FP- False Positive

FPR- False Positive Rate

FRP-Fiberglass Reinforced Pipe

Gi-Gini Index

GIS- Geographic Information System

HDPE-High-Density Polyethylene

IIMM- International Infrastructure Management Manual

k-NN-k-Nearest Neighbors

LR- Logistic Regression

MDA-Mean Decrease Accuracy

MDI-Mean Decrease Impurity

ML-Machine Learning

MLR-Multinomial Logistic Regression

MSSAM-Municipal Sewage System Asset Management

MS4-Municipal Separate Storm Sewer System

NASSCO- National Association of Sewer Service Company

NPDES-National Pollutant Discharge Elimination System

OTHERS- Pipes with unknown material

PCCP-Pre-stressed Concrete Cylinder Pipe

PACP- Pipeline Assessment Certification Program

PVC- Polyvinyl Chloride

RCP-Reinforced Concrete Pipe

RF- Random Forests

RMP-Reinforced Polymer Mortar

ROC- Receiver Operator Characteristic

SCRAPS-Sewer Cataloging, Retrieval and Prioritization System

SSS-Storm Sewer Systems

SVM- Support Vector Machine

SMOTE-Synthetic Minority Oversampling Technique

TN-True Negative

TP-True Positive

TPR-True Positive Rate

UnReinCONC-Unreinforced Concrete Pipe

Upstream MH-Upstream Manhole

VCP-Vitrified Clay Pipe

WRc-Water Research Center