Predicting Inspection Priority and Pipe Breaks in Water Distribution Network using Machine Learning

Tanay Kulkarni[1], Devashri Karve[2], Yijie Zhu[3], and Zulkifli Palinrungi[4]

1 Graduate Student, Department of Civil and Environmental Engineering, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213. Email: tskulkar@andrew.cmu.edu

2 Graduate Student, Department of Civil and Environmental Engineering, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213. Email: dkarve@andrew.cmu.edu

3 Graduate Student, Department of Civil and Environmental Engineering, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213. Email: yijiezhu@andrew.cmu.edu

4 Graduate Student, Department of Energy Science, Technology, and Policy, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213. Email: zpalinru@andrew.cmu.edu

ABSTRACT

Every year, water utilities worldwide spend millions of dollars to plan, design, build, operate, and maintain their water network infrastructure systems. The water distribution networks play a crucial role in ensuring clean potable water to millions of people and are vital to the entire infrastructure system. As years go by, these assets age, deteriorate, and fail, causing service interruptions and system-wide disruptions. Often a blend of geographic, physical, and hydraulic parameters impacts this deterioration. The objective of this study is to predict the condition of pipelines considering the geophysical parameters such as soil type, groundwater table, diameter, material, length, and hydraulic data such as pressure, velocity, discharge, and year of installation, and recommendations for pipe inspection priorities as well as prediction of future pipe breaks. The study proposes a machine learning models to predict pipe inspection priorities and the number of pipe breaks.

INTRODUCTION, BACKGROUND, AND NECESSITY FOR PROJECT

In general, the clean water supply goes through several stages, starting from tapping water from a source and then disinfecting and purifying it. After the purification process, the water is stored in local reservoirs before supplying it to customers. The wastewater generated after the citizens consume water is then disposed back into the source after treating the sanitary wastewater [1].

With population growth and urban development, the water infrastructure becomes more complex, particularly in determining the effectiveness of improvement investments [2]. Due to funding limitations, the determination of repairs must be based on priorities, which will measure how much impact will be generated from each improvement that will be made. Uniquely, water infrastructure is majorly laid underground and cannot be seen, which is very different from roads and bridges that can be visually inspected. With such an extended infrastructure with different characteristics, prioritization is very difficult if it is not inspected to assess the condition of the infrastructure [2].

Pipe inspection by digging all water infrastructure installations is not possible due to high costs and ineffective work. To address this complex problem, this study tries to apply several methods to generate pipeline inspection priorities to make the inspection work more effective and significantly impact the water infrastructure system.

Based on the ASCE infrastructure report card, the grade of water infrastructure in America gets C- [3]. Most of the water infrastructure in America is aging, and out of the 2.2 million miles of water pipes, around 12,000 miles have been replaced in 2020 alone. The report also explains that there is a pipe break in every 2 minutes, and about 6 trillion gallons are lost every day [3].

The infrastructure design in many places is designed based on the assumption of population growth or static without any future design scenarios that allow for unexpected developments. After such extensive and long-running development, much of the water infrastructure is aging and deteriorating [2]. This damage is caused by many factors such as soil type, groundwater table, diameter, material, length, and hydraulic data such as pressure, velocity, discharge, and year of installation [4].

This study uses water infrastructure data from Watertown, Connecticut, as a representative to obtain the right analysis method to determine the priority of pipe inspections and projected leaks in each of the installed pipe. Machine learning algorithms are used to combine existing variables into several more comprehensive damage parameters in this research. This research deploys decision tree and random forest classification to determine the inspection priority. Meanwhile, to predict pipe leakage, decision tree and random forest regression are used. By knowing the prediction of future pipe leaks, water utility companies can reduce capital losses by intelligently replacing pipes before the breaks occur. This study compares the accuracy of the decision tree method and random forest both in priority and break predictions.

DATA COLLECTION

The infrastructure management plays a crucial role in water infrastructure systems. A good condition water infrastructure will provide potable water to a city with millions of people, sometimes even beyond the design life of these assets. These systems, that are affected by many factors like deterioration, fail causing service interruptions, water quality issues, etc.

The water infrastructure data for the city of Watertown, Connecticut was obtained from Bentley Systems Inc. The data included but was not limited to the pipeline network characteristics data viz., the historical breaks reported between 2015 and 2020, diameter, year of installation, material, maximum pressure, maximum discharge, soil pH, and groundwater depth.

WATERTOWN WATER NETWORK INFRASTRUCTURE

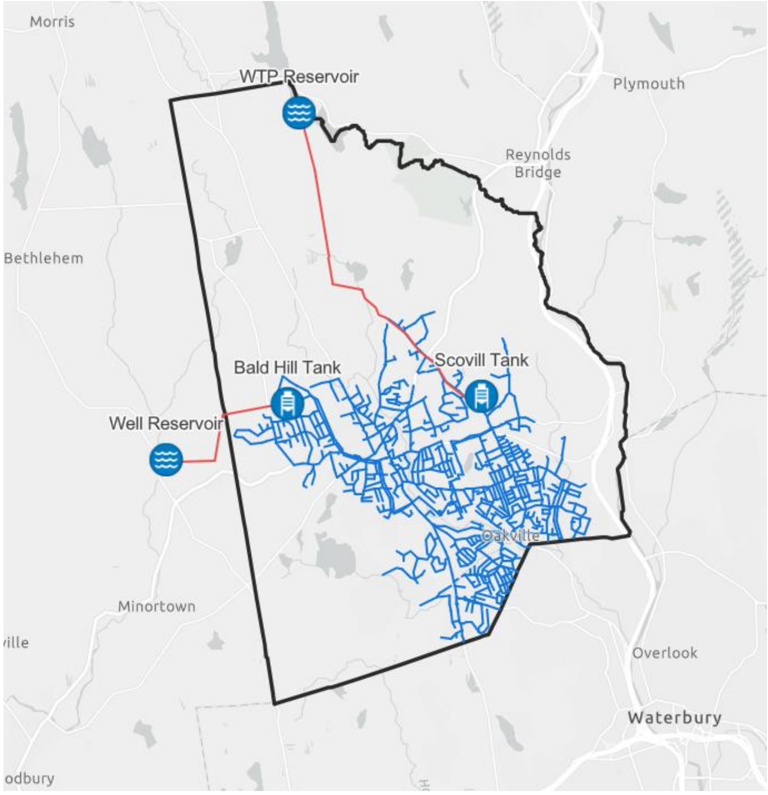


Figure 1 GIS Map of Water Distribution Network Infrastructure System, Watertown, CT

The total pipeline network length in Watertown is approximately, 110 miles. The pipe age ranges between 40 to 112 years. The network is sub-divided into 6 operational zones that caters to a total population of 22,000 in 2020. The system has around 6610 customer meters installed throughout the city. Figure 1 shows a GIS map of Watertown's water distribution network infrastructure.

METHODOLOGY

The methodology adopted in this study is presented in the Figure 2. The collected data is cleaned for outliers, missing data, and checked for anomalies. An exploratory data analysis is performed over the data to understand the distribution of the independent random variables and checked for necessary transformations.

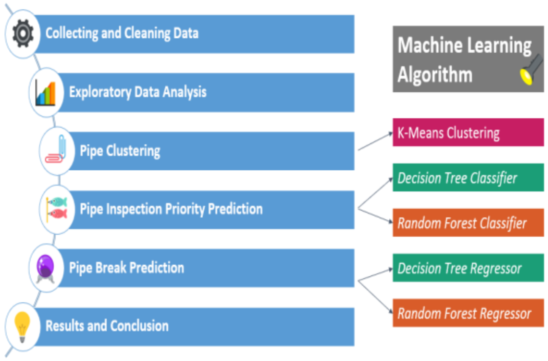


Figure 2 Project Methodology

The pipes are then clustered using K-Means Clustering to identify pipes having similar traits. Based on clusters identified, pipes are classified for inspection priorities as ‘high’, ‘medium’, and ‘low’. A machine learning model is trained and tested to predict pipe inspection priority based on Decision Tree and Random Forest Classifiers. A machine learning model is trained and tested to predict the number of breaks based on Decision Tree and Random Forest Regressors. The results are presented, and the accuracy of the ML models is discussed.

EXPLORATORY DATA ANALYSIS

Figure 3 shows the pipes on which the customers had reported breaks.

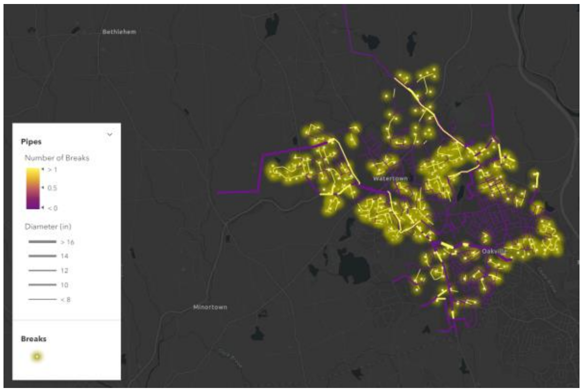


Figure 3 Map: Location of Reported Breaks between 2019 and 2020

Figure 4 shows that pipes installed between 1946 and 1981 have a greater number of breaks reported as compared to the pipes that were installed before 1946.

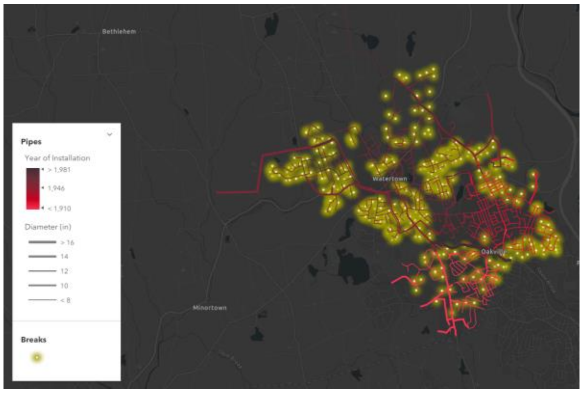


Figure 4 Map: Geospatial Correlation between Breaks and Installation Year

Figure 5 shows that there is a strong correlation between pipe material and breaks. Majority of the pipe breaks occurred on Ductile Iron pipes.

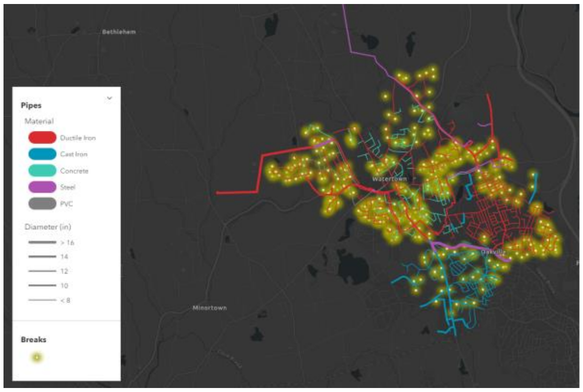


Figure 5 Map: Geospatial Correlation between Breaks and Material

Figure 6 shows that there is a strong "positive" correlation between pressure and number of breaks on pipes. That means higher the pressure on smaller diameter pipes (insight from previous map) a greater number of breaks there will be.

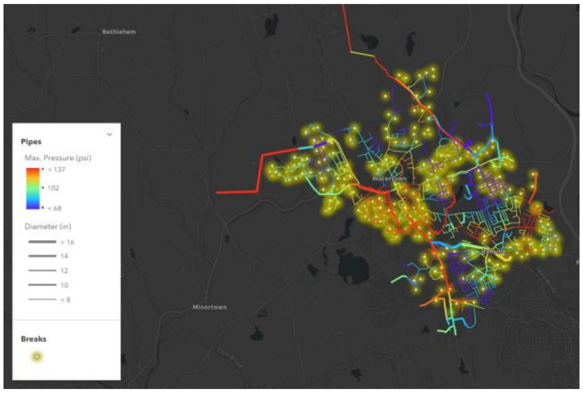


Figure 6 Map: Geospatial Correlation between Breaks and Pressure

Figure 7 shows that there is a strong "negative" correlation between discharge and number of breaks reported. That means, a greater number of breaks were reported on pipes with lower discharges.

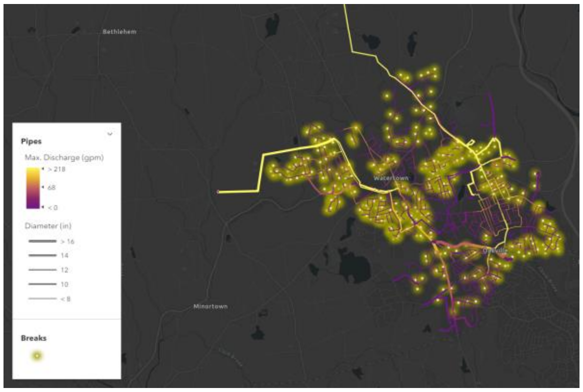


Figure 7 Map: Geospatial Correlation between Breaks and Discharge

Figure 8 shows a strong "negative" correlation between bed-soil pH and number of pipe breaks. This indicates that a greater number of breaks were observed on pipes that were laid in soils with lower pH values (approx. less than pH of 5.3).

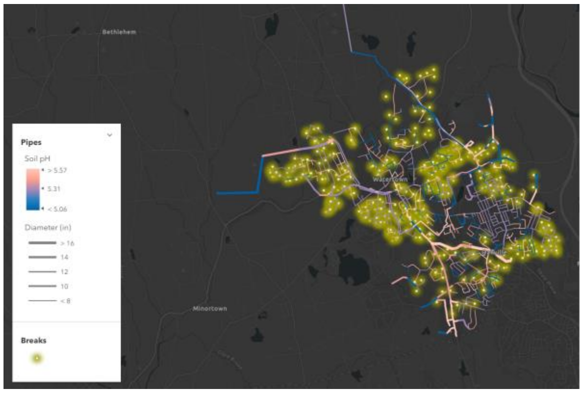


Figure 8 Map: Geospatial Correlation between Breaks and Bed-soil pH

Ironically, Figure 9 shows that there is no visible correlation between number of breaks and depth of groundwater table.

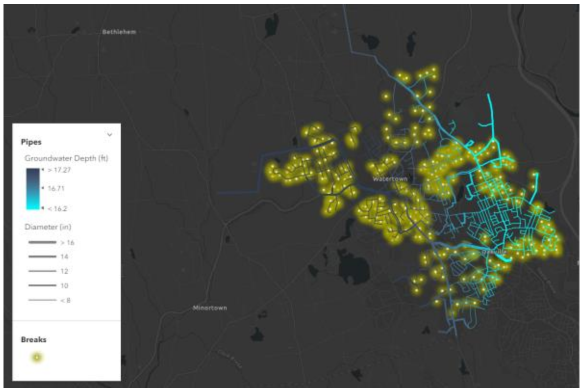


Figure 9 Map: Geospatial Correlation between Breaks and Groundwater Depth

Figure 10 shows that maximum number of pipes are older than 70 years and are of Ductile Iron pipes.

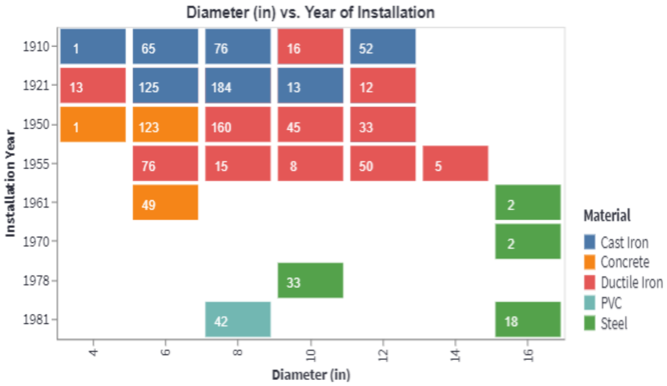


Figure 10 Diameter vs. Installation Year

Figure 11 shows that older ductile iron pipes have maximum records of breaks.

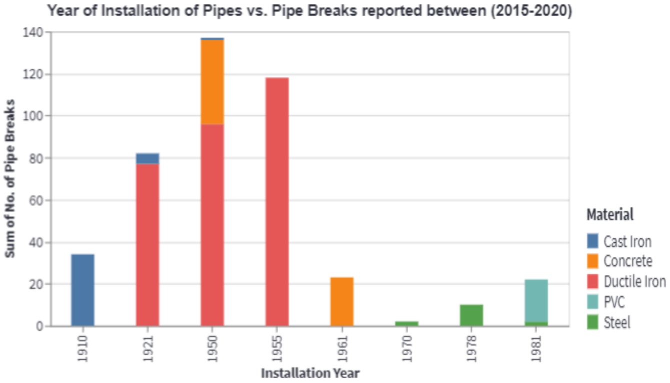


Figure 11 Installation Year vs. Total Breaks

Figure 12 shows that older pipes carry maximum discharge in the network. Should anything happen to these pipes, maximum number of customers would be effectively impacted.

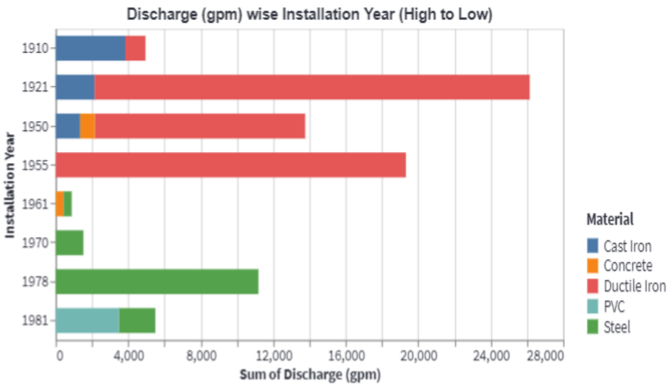


Figure 12 Discharge vs. Installation Year

Figure 13 shows that pH < 6 results in higher corrosion and the number of breaks reported are more in pipes with bed-soil pH less than 6.

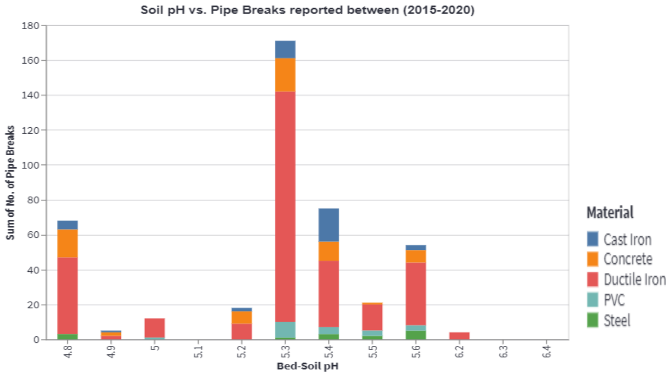


Figure 13 Bed-soil pH vs. Total Breaks

DATA SCIENCE PROCEDURES

The raw dataframe contains attribute fields viz. ‘Diameter’, ‘Length’, ‘Material’, ‘Groundwater Depth’, ‘Installation Year’, ‘Age’, ‘Pressure’, ‘Discharge’, ‘Soil pH’, and ‘Number of Breaks’. The dataframe is randomly split in 75% ‘Model’ dataset and 25% ‘Test’ dataset. The Model dataset is further cleaned, and the outliers are handled by snapping the outliers to the 25th and 75th percentile inter-quartile range on lower and upper bounds, respectively.

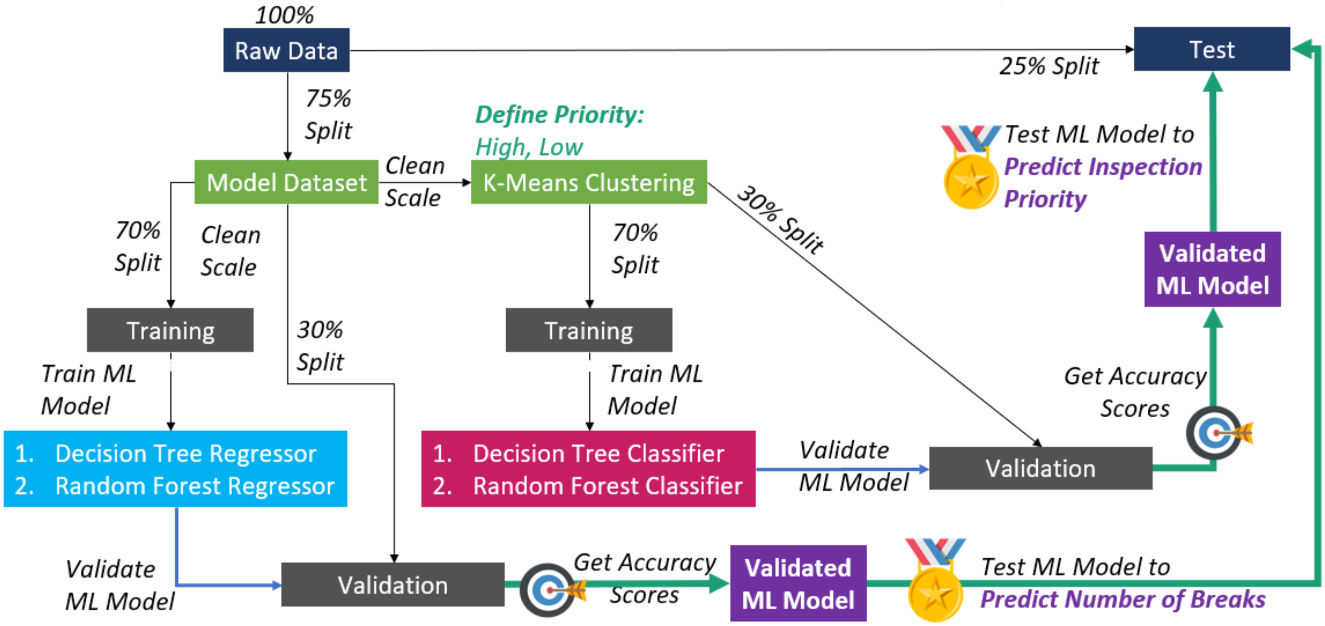


Figure 14 Project Methodology

FOR PIPE INSPECTION PRIORITY PREDICTION

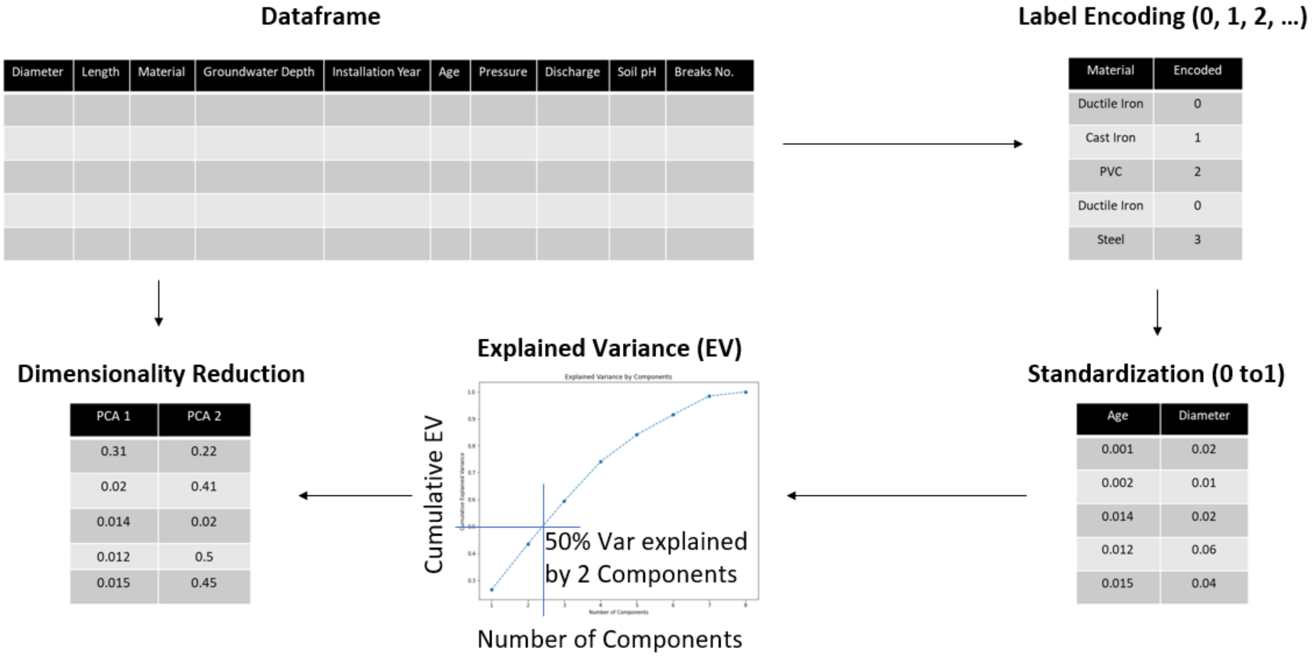
K-Means clustering is performed on the model dataset. Based on the cluster insights, the pipes are classified as ‘High’ and ‘Low’ inspection priorities. This is discussed in the next section in more detail. The clustered model dataset is further randomly split into 70% training and 30% validation datasets. The training dataset is used to train the data by Decision Tree and Random Forest Classifier. The trained ML model is then validated against the validation dataset. Accuracy scores are checked and then the validated ML model is applied on the ‘Test’ dataset to predict the pipe inspection priorities.

FOR PIPE BREAK PREDICTION

Similarly, the ‘Model’ dataset is randomly split into 70% training and 30% validation datasets. The training dataset is used to train the data by Decision Tree and Random Forest Regressors. The trained ML model is then validated against the validation dataset. Accuracy scores are checked and then the validated ML model is applied on the ‘Test’ dataset to predict the number of pipe breaks.

K-MEANS CLUSTERING

K-Means clustering is one of the most popular unsupervised machine learning algorithms. It is used for solving classification problems. K-Means segregates the unlabeled data into various groups, called as clusters, based on having similar features and patterns. It is a distance-based algorithm and therefore all the variables are scaled between 0 and 1.



Dataframe

Diameter	Length	Material	Groundwater Depth	Installation Year	Age	Pressure	Discharge	Soil pH	Breaks No.

Label Encoding (0, 1, 2, …)

Material	Encoded
Ductile Iron	0
Cast Iron	1
PVC	2
Ductile Iron	0
Steel	3

Standardization (0 to1)

Age	Diameter
0.001	0.02
0.002	0.01
0.014	0.02
0.012	0.06
0.015	0.04

Dimensionality Reduction

PCA 1	PCA 2
0.31	0.22
0.02	0.41
0.014	0.02
0.012	0.5
0.015	0.45

Figure 15 Dimensionality Reduction

The dataframe consists of ‘Material’ field which is a non-numerical field. Therefore, unique labels of material are encoded as integers so that it can be used in clustering. All the fields are then scaled between 0 and 1 to ensure that it follows standardization. Principal Component Analysis (PCA) is a widely used method for dimensionality reduction of data. Initially the dataframe consisted of 10 fields which upon PCA were reduced to 2 components. The number of components is fixed based on the cumulative explained variance of 50%.

The number of optimal clusters for using K-Means clustering can be determined using the Elbow method and is cross-checked using the Silhouette scores. The Elbow graph showing number of clusters versus the sum of squared errors is shown in Figure 16 and the silhouette scores are shown in Figure 17. Based on the above analysis it is determined that the optimal number of clusters is 3.

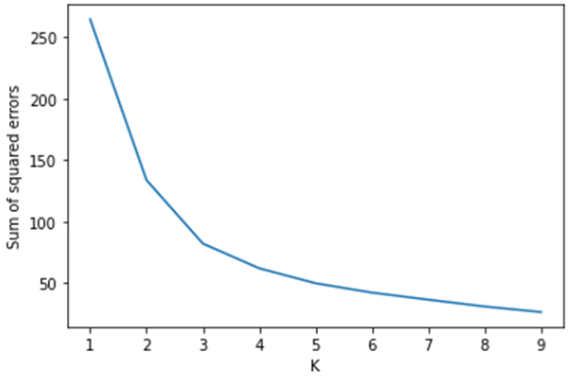
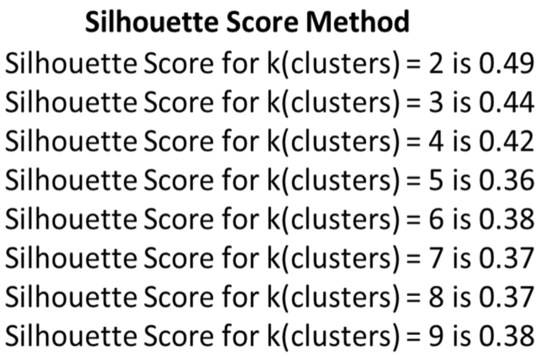


Figure 16 Elbow Method to Determine Optimal Number of Clusters



Silhouette Score Method

Silhouette Score for k(clusters) = 2 is 0.49
Silhouette Score for k(clusters) = 3 is 0.44
Silhouette Score for k(clusters) = 4 is 0.42
Silhouette Score for k(clusters) = 5 is 0.36
Silhouette Score for k(clusters) = 6 is 0.38
Silhouette Score for k(clusters) = 7 is 0.37
Silhouette Score for k(clusters) = 8 is 0.37
Silhouette Score for k(clusters) = 9 is 0.38

Figure 17 Silhouette Scores to Determine Optimal Number of Clusters

After performing K-Means clustering on the dataframe, the pipes are clustered as 0, 1, and 2, i.e., in three clusters. A histogram of clusters is shown in Figure 18. Clusters 0, 1, and 2 have 340, 252, and 322 number of unique pipes within them.

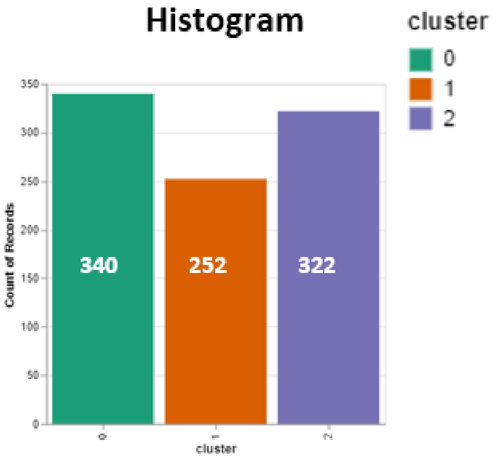


Figure 18 Cluster Histogram

The chart in Figure 19 shows the box plots with outliers of different variables for each cluster. The line represents the mean (average) values. From Figure it can be observed that, cluster 0 has relatively pipes of lowest diameters, highest length, highest breaks, medium age, highest number of customers, lowest bed-soil pH, medium pressure, and lower discharges. These are the pipes that represent the tertiary segments of water distribution networks that connects the primary and secondary mains to the customers.

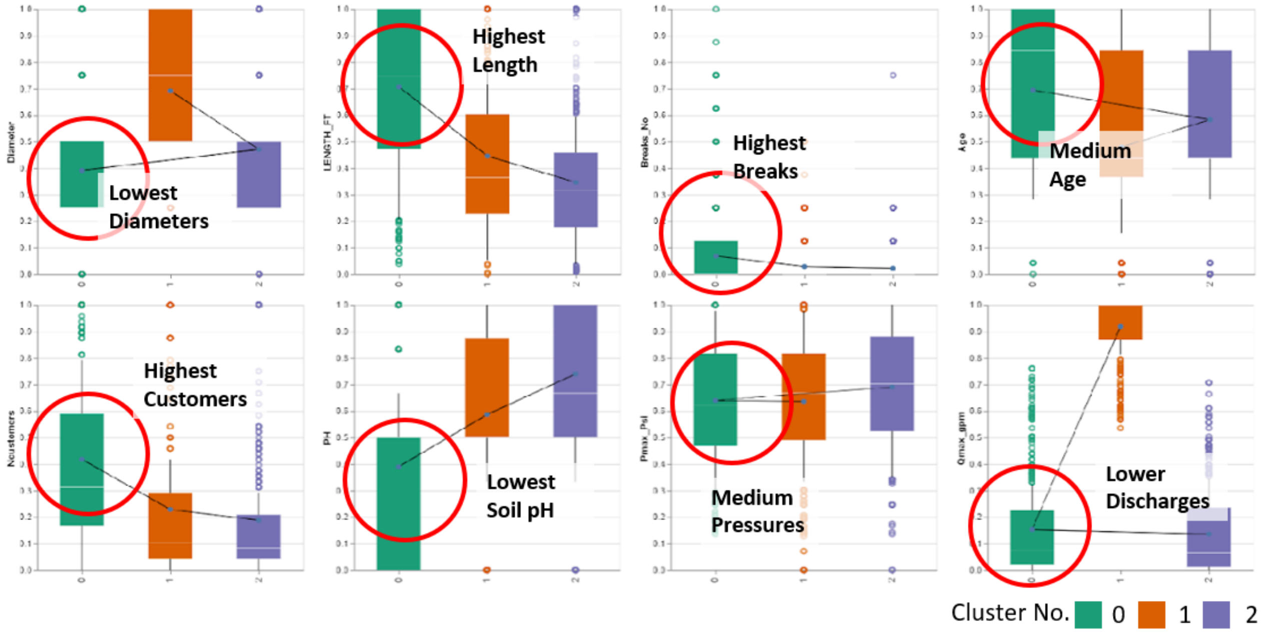


Figure 19 K-Means Cluster Boxplots with Outliers

It must also be noted that should these pipes (cluster 0) fail, then there would be largest system wide disruption, and highest number of customers will be directly impacted. Based on this understanding, Cluster 0 is considered to have highest inspection priority, and cluster 1 and 2 are considered to have relatively 'lower' inspection priorities. The inspection priorities are defined in such a way that it can either be high or low.

PREDICTING PIPE INSPECTION PRIORITIES

Prediction of inspection priorities is categorized as a classification problem. The objective here is to train 70% of the clustered model dataset using Decision Tree and Random Forest Classifiers and then validate the predictions on the remaining 30% of the clustered model dataset. Figure 20 explains the methodology adopted here.

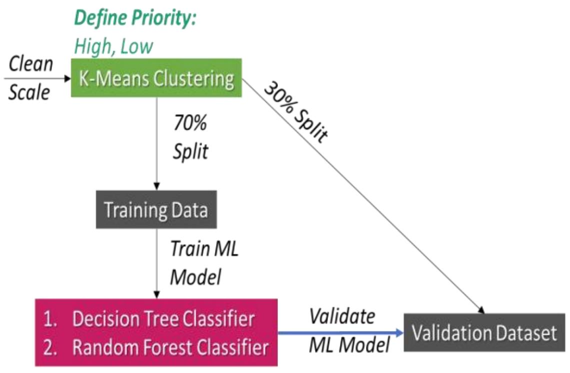


Figure 20 Training Machine Learning Classifiers

PREDICTING NUMBER OF BREAKS

Prediction of number of breaks is categorized as a regression problem. The objective here is to train 70% of the ‘Model’ dataset using Decision Tree and Random Forest Regressors and then validate the predictions on the remaining 30% of the clustered model dataset. Figure 21 explains the methodology adopted here.

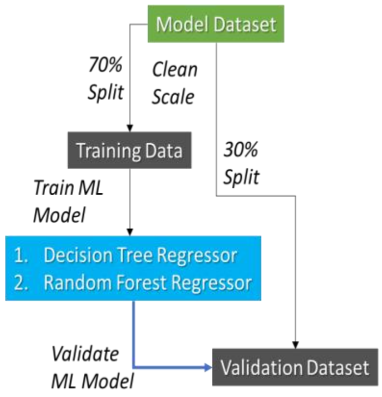


Figure 21 Training Machine Learning Regressors

RESULTS AND DISCUSSION

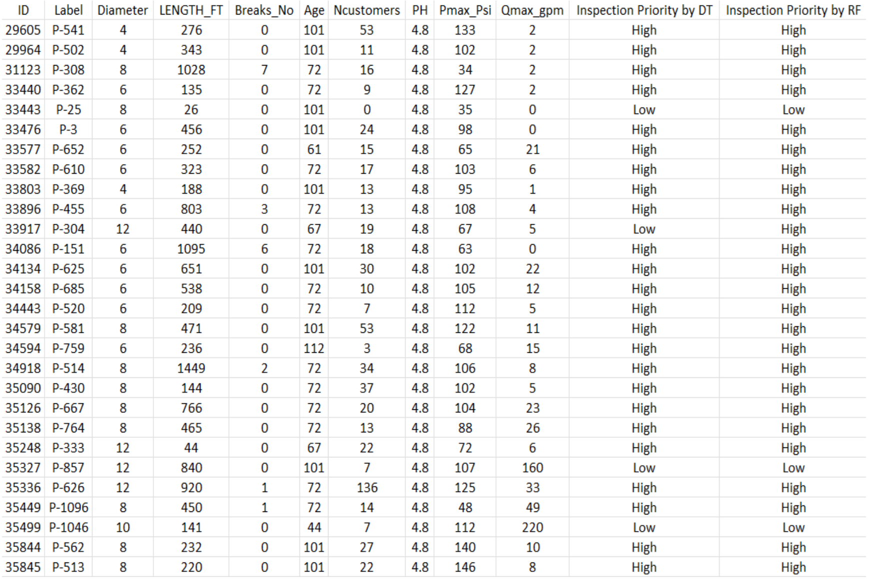
PREDICTING INSPECTION PRIORITIES

Upon training the Classification ML models for predicting the inspection priorities, the accuracy of the two classifiers on the validation data is shown in Table 1.

Table 1 ML Validation Results- Classification

	Decision Tree Classification	Random Forest Classification
Mean Squared Error	0.069	0.112
Accuracy	93.29%	88.72%
Confusion Matrix	[87 15] [4 169]	[84 18] [13 160]

The Decision Tree with 93.29% accuracy performed better than Random Forest with an accuracy of 88.72%. The validated ML model is then test on the 25% of the 'Test' dataset and the inspection priorities were predicted successfully both by Decision Tree and Random Forest Classifier. Figure 22 shows random samples from the 'Test' dataset with predicted inspection priorities.



ID	Label	Diameter	LENGTH_FT	Breaks_No	Age	Ncustomers	PH	Pmax_Psi	Qmax_gpm	Inspection Priority by DT	Inspection Priority by RF
29605	P-541	4	276	0	101	53	4.8	133	2	High	High
29964	P-502	4	343	0	101	11	4.8	102	2	High	High
31123	P-308	8	1028	7	72	16	4.8	34	2	High	High
33440	P-362	6	135	0	72	9	4.8	127	2	High	High
33443	P-25	8	26	0	101	0	4.8	35	0	Low	Low
33476	P-3	6	456	0	101	24	4.8	98	0	High	High
33577	P-652	6	252	0	61	15	4.8	65	21	High	High
33582	P-610	6	323	0	72	17	4.8	103	6	High	High
33803	P-369	4	188	0	101	13	4.8	95	1	High	High
33896	P-455	6	803	3	72	13	4.8	108	4	High	High
33917	P-304	12	440	0	67	19	4.8	67	5	Low	High
34086	P-151	6	1095	6	72	18	4.8	63	0	High	High
34134	P-625	6	651	0	101	30	4.8	102	22	High	High
34158	P-685	6	538	0	72	10	4.8	105	12	High	High
34443	P-520	6	209	0	72	7	4.8	112	5	High	High
34579	P-581	8	471	0	101	53	4.8	122	11	High	High
34594	P-759	6	236	0	112	3	4.8	68	15	High	High
34918	P-514	8	1449	2	72	34	4.8	106	8	High	High
35090	P-430	8	144	0	72	37	4.8	102	5	High	High
35126	P-667	8	766	0	72	20	4.8	104	23	High	High
35138	P-764	8	465	0	72	13	4.8	88	26	High	High
35248	P-333	12	44	0	67	22	4.8	72	6	High	High
35327	P-857	12	840	0	101	7	4.8	107	160	Low	Low
35336	P-626	12	920	1	72	136	4.8	125	33	High	High
35449	P-1096	8	450	1	72	14	4.8	48	49	High	High
35499	P-1046	10	141	0	44	7	4.8	112	220	Low	Low
35844	P-562	8	232	0	101	27	4.8	140	10	High	High
35845	P-513	8	220	0	101	22	4.8	146	8	High	High

Figure 22 *Random Samples from the 'Test' Dataset with Predicted Inspection Priorities*

Figure 23 and Figure 24 shows maps of 25% Test data with predicted inspection priorities using Decision Tree and Random Forest Classifiers, respectively.

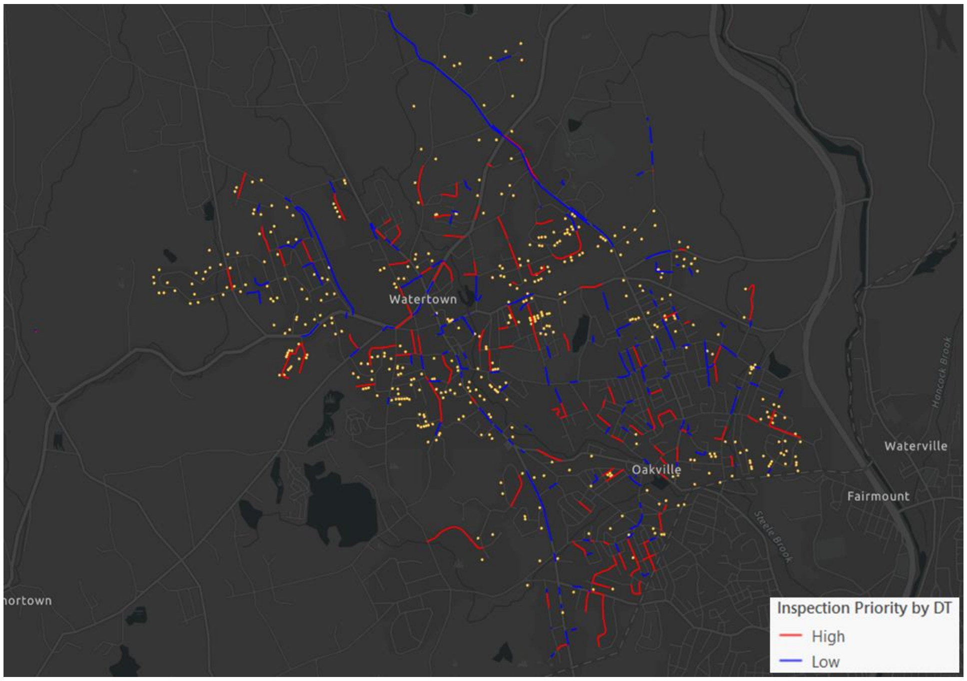


Figure 23 *Map: Decision Tree Model's Inspection Priority Predictions on 25% Test Data*, Accuracy: 93.29%

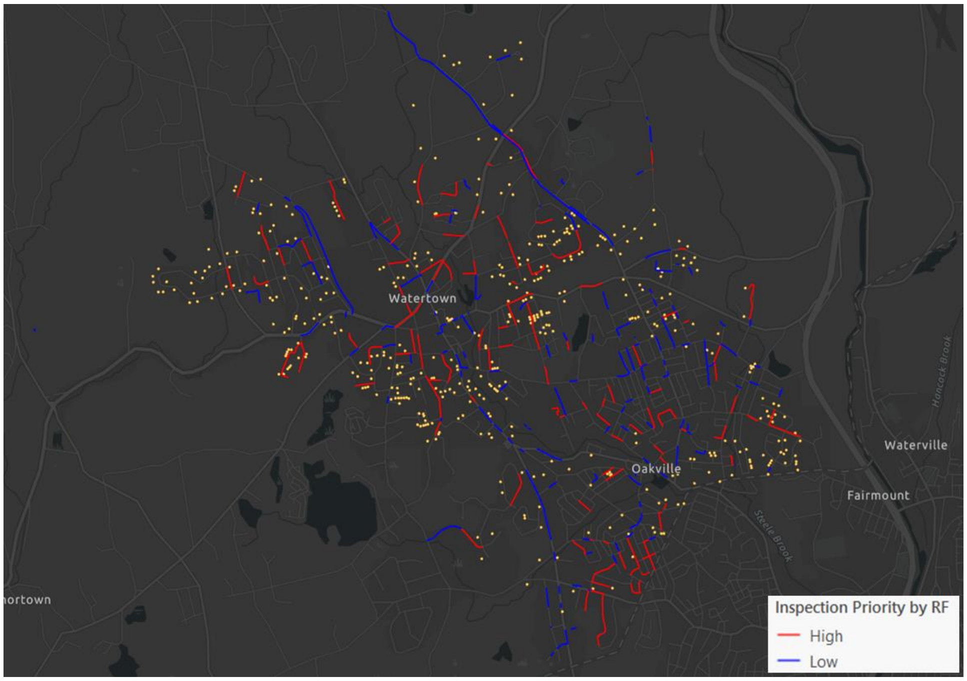


Figure 24 *Map: Random Forest Model's Inspection Priority Predictions on 25% Test Data, Accuracy: 88.72%*

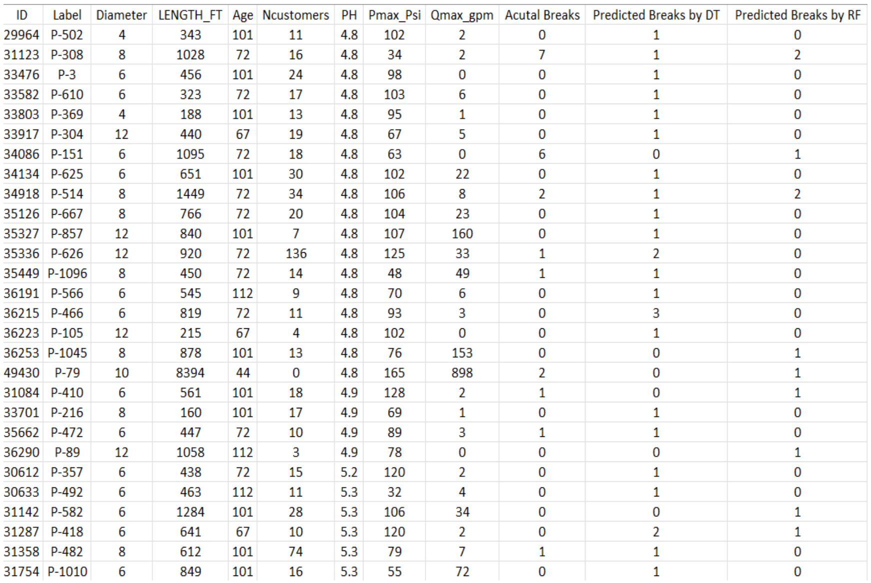
PREDICTING NUMBER OF BREAKS

Upon training the Regression ML models for predicting the number of breaks, the accuracy of the two regressors on the validation data is shown in Table 2.

Table 2 ML Validation Results- Regression

	Random Forest Regression	Decision Tree Regression
Mean Squared Error	0.93	1.16
Accuracy	78.90%	71.27%
Confusion Matrix	[207 19 0 0 0 0 0] [22 8 1 0 0 0 0] [10 2 2 0 0 0 0] [1 0 0 0 0 0 0] [0 1 0 0 0 0 0] [1 0 0 0 0 0 0] [0 1 0 0 0 0 0]	[184 26 9 3 4 0 0] [15 12 1 1 0 2 0] [10 2 0 2 0 0 0] [1 0 0 0 0 0 0] [0 1 0 0 0 0 0] [0 0 0 0 0 0 0] [1 0 0 0 0 0 0]

The Random Forest with 78.90% accuracy performed better than Decision Tree with an accuracy of 71.27%. The validated ML model is then test on the 25% of the ‘Test’ dataset and the number of breaks were predicted successfully both by Decision Tree and Random Forest Regressors. Figure 25 shows random samples from the ‘Test’ dataset with predicted number of breaks.



ID	Label	Diameter	LENGTH_FT	Age	Ncustomers	PH	Pmax_Psi	Qmax_gpm	Acutal Breaks	Predicted Breaks by DT	Predicted Breaks by RF
29964	P-502	4	343	101	11	4.8	102	2	0	1	0
31123	P-308	8	1028	72	16	4.8	34	2	7	1	2
33476	P-3	6	456	101	24	4.8	98	0	0	1	0
33582	P-610	6	323	72	17	4.8	103	6	0	1	0
33803	P-369	4	188	101	13	4.8	95	1	0	1	0
33917	P-304	12	440	67	19	4.8	67	5	0	1	0
34086	P-151	6	1095	72	18	4.8	63	0	6	0	1
34134	P-625	6	651	101	30	4.8	102	22	0	1	0
34918	P-514	8	1449	72	34	4.8	106	8	2	1	2
35126	P-667	8	766	72	20	4.8	104	23	0	1	0
35327	P-857	12	840	101	7	4.8	107	160	0	1	0
35336	P-626	12	920	72	136	4.8	125	33	1	2	0
35449	P-1096	8	450	72	14	4.8	48	49	1	1	0
36191	P-566	6	545	112	9	4.8	70	6	0	1	0
36215	P-466	6	819	72	11	4.8	93	3	0	3	0
36223	P-105	12	215	67	4	4.8	102	0	0	1	0
36253	P-1045	8	878	101	13	4.8	76	153	0	0	1
49430	P-79	10	8394	44	0	4.8	165	898	2	0	1
31084	P-410	6	561	101	18	4.9	128	2	1	0	1
33701	P-216	8	160	101	17	4.9	69	1	0	1	0
35662	P-472	6	447	72	10	4.9	89	3	1	1	0
36290	P-89	12	1058	112	3	4.9	78	0	0	0	1
30612	P-357	6	438	72	15	5.2	120	2	0	1	0
30633	P-492	6	463	112	11	5.3	32	4	0	1	0
31142	P-582	6	1284	101	28	5.3	106	34	0	0	1
31287	P-418	6	641	67	10	5.3	120	2	0	2	1
31358	P-482	8	612	101	74	5.3	79	7	1	1	0
31754	P-1010	6	849	101	16	5.3	55	72	0	1	0

Figure 25 *Random Samples from the 'Test' Dataset with Predicted Number of Breaks*

Figure 26 and Figure 27 shows maps of 25% Test data with predicted number of breaks using Decision Tree and Random Forest Regressors, respectively.

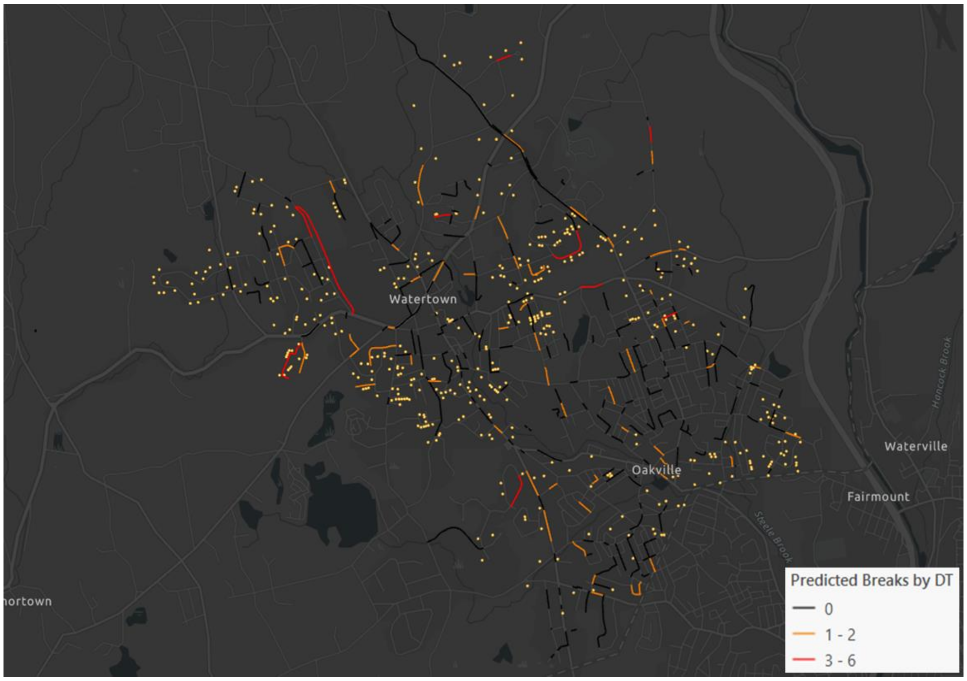


Figure 26 *Map: Decision Tree Model's Number of Breaks Predictions on 25% Test Data, Accuracy: 71.27%*

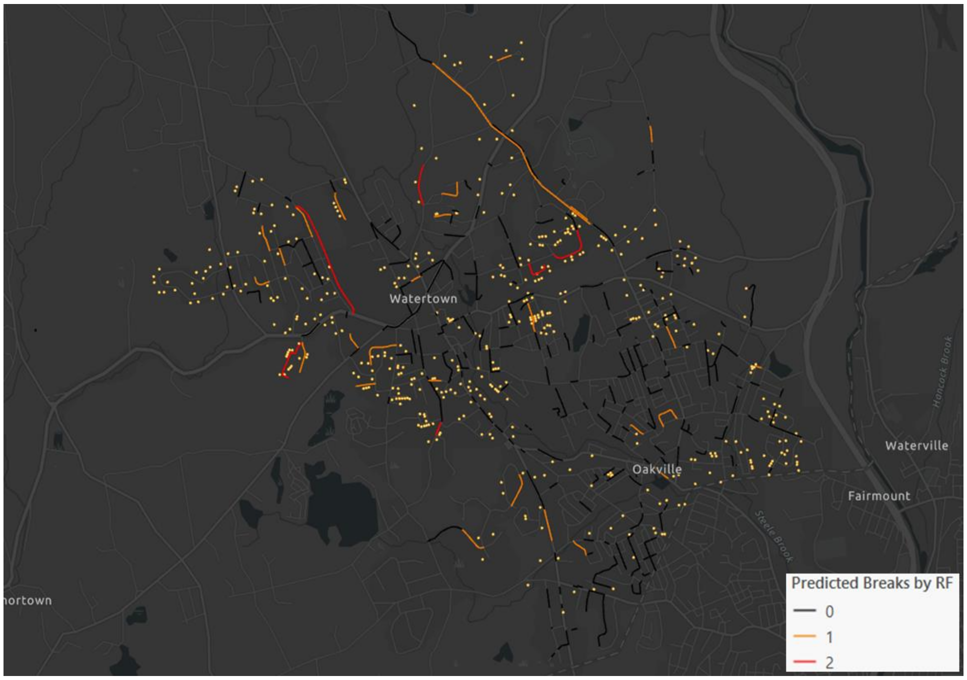


Figure 27 *Map: Random Forest Model's Number of Breaks Predictions on 25% Test Data, Accuracy: 78.90%*

CONCLUSIONS

The objective of this study to predict the condition of pipelines considering the geophysical parameters such as soil type, groundwater table, diameter, material, length, and hydraulic data such as pressure, velocity, discharge, and year of installation, and recommendations for pipe inspection priorities as well as prediction of future pipe breaks, is achieved successfully. It is evident that machine learning has a great scope for applications in operations and maintenance of water infrastructure systems. The methodology proposed in this study proved to result in better accuracy in predictions. Decision Tree (accuracy 93.29%) performed better than Random Forest (accuracy 88.73%) to handle the Classification problem to predict the inspection priorities of pipes. Random Forest (accuracy 78.90%) performed better than Decision Tree (accuracy 71.27%) to handle the Regression problem to predict the number of breaks.

The scope of this project can be extended to planning of maintenance activities using predictions and asset-to-asset relation in the future. It can also be extended to considering influence of critical assets of water infrastructure over the critical assets of transportation infrastructure, and vice versa. Understand influence of dynamic features such as pressure, discharge, age, etc. that change over time plays a critical role in the overall lifecycle of assets. Consideration of such temporally dynamic features to build a dynamic prediction model can be explored in the future.

ACKNOWLEDGEMENTS

This work is a part of the 12-750 Infrastructure Management course at Carnegie Mellon University under the guidance of Prof. Dr. Donald Coffelt and Prof. Dr. Qiling Zou. The authors wish to thank Dr. Coffelt for his guidance, advice, support, and motivation. The authors also would like to thank Dr. Thomas Walski and Shar Govindan from Bentley Systems Inc. for the invaluable support, mentorship. The sample dataset of Watertown is provided by Bentley Systems Inc. The authors would also like to thank the teaching assistants Dr. Jiachao Liu and Kevin Freymiller for their feedback and suggestions.

REFERENCES

1. “Drinking Water,” ASCE’s 2021 Infrastructure Report Card I, Jan. 17, 2017. https://infrastructurereportcard.org/cat-item/drinking-water/ (accessed May 04, 2022)
2. “Water Network Infrastructure Services I DTK Hydronet Solutions,” DTK Hydronet Solutions. https://www.dtkhydronet.com/water (accessed May 04, 2022)
3. F. Chughtai and T. Zayed, “Infrastructure Condition Prediction Models for Sustainable Sewer Pipelines,” Journal of Performance of Constructed Facilities, vol. 22, no. 5, pp. 333–341, Oct. 2008, doi: 10.1061/(ASCE)0887-3828(2008)22:5(333)
4. J. B. Hollander, K. Pallagst, T. Schwarz, and F. J. Popper, “Planning shrinking cities,” Progress in planning, vol. 72, no. 4, pp. 223–232, 2009