Joint CSCE Construction Specialty & CRC Conference 2025
Conférence conjointe spécialisée en construction de la SCGC et CRC-2025

Montreal, Quebec
July 28-31, 2025 / *28-31 juillet 2025*

IDENTIFYING LEAD SERVICE LINES USING DATA MINING TECHNIQUES

Abuawad, I.[1], Azarian, A.[2], Elhosary, E.[2], Hammam, M.[2], Hedaiaty Marzouny, N.[2], and Nik-Bakht, M.[3]

[1] PhD Student, Department of Construction Engineering, École de Technologie Supérieure ÉTS, Montreal, Canada
[2] PhD Student, Dept. of Building, Civil, and Environmental Engineering, Concordia University, Montreal, Canada
[3] Associate Professor, Dept. of Building, Civil, and Environmental Engineering, Concordia University, Montreal, Canada

ABSTRACT: The water distribution system supplies drinking water to households through service lines (SLs), which may contain lead, posing significant health risks, especially for children. Due to its durability and corrosion resistance, lead was widely used in residential service lines in the U.S., leaving many municipalities uncertain about the number and locations of remaining Lead Service Lines (LSLs). With updated health regulations and growing public concern, municipalities must replace LSLs, but challenges such as high replacement costs, complex tap water testing, and incomplete pipe inventories hinder efforts. This study addresses these uncertainties by applying a data mining approach to predict LSL locations. A DBSCAN clustering algorithm was used to identify priority areas, followed by the development and evaluation of three predictive models: Decision Tree (DT), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). KNN demonstrated strong performance with an F1-score of 99.0% in two out of three scenarios, while DT outperformed in one scenario with an F1-score of 99.8%. Given its consistency and high recall, KNN was integrated into a decision-making tool to help the municipality of Aurora, Illinois efficiently and accurately predict LSLs. This study provides a scalable framework for municipalities to identify and replace lead pipes, improving public health and resource allocation.

1. INTRODUCTION

The water distribution system (WDS) is a critical urban infrastructure system that supplies households with drinking water (Berglund et al., 2023). The lateral or service line (SL) is the small diameter pipe that connects buildings to the larger water mains. Since 1900, due to its durability and corrosion resistance, lead has dominated the materials used for water lateral in the US residential buildings (Hajiseyedjavadi et al., 2020). In 2014, lead water contamination gained attention all over the US, followed by the Flint crisis of drinking water (Gurewitsch et al., 2017; Pauli, 2020). Since then, the problem has been a national public health concern that necessitates interventions on all community scales (Zartarian et al., 2022).Besides water, paints and soil are the other primary sources of lead exposure risk (Gould, 2009). Lead exposure has reported health consequences, especially for children, it is associated with cognitive and behavioral impairment stemming from biological and neurologic damage (Bellinger, 2008b, 2008a). In drinking water, Lead service lines (LSLs) are the major source of lead accounting for 50% to 86% of total lead measured in tap water (Brown & Margolis, 2012; Hensley et al., 2021; Lytle et al., 2021).

In 1986, the US started a regulation revision that resulted in a lead ban from being used in plumbing (Hajiseyedjavadi et al., 2020). In 2021, the U.S. Environmental Protection Agency (USEPA) in its revision established a threshold lead action level of 10 ppb concentration in tap water (Hensley et al., 2021). Municipalities were required to monitor lead concentration in tap water and apply control measures. When control measures are ineffective, municipalities are mandated to replace lead pipes. Despite efforts since the 1970s to reduce lead exposure from multiple sources, children still show elevated blood lead levels (BLL) (Brown & Margolis, 2012), partly due to the continued use of lead service lines (LSLs) and the challenges limiting their replacement. Because replacement plans are constrained by many challenges, LSLs are still serving many houses (Hajiseyedjavadi et al., 2020).

On the other hand, the advancement in information and communication technology (ICT) aligned with open data policies paved the way for digitalization. Data mining is a powerful digital tool that enhances decision-making in water infrastructure management, addressing limitations of conventional systems and improving outcomes. This study utilizes data mining techniques, attempting to reduce the uncertainties in predicting lead presence in service lines for the city of Aurora, Illinois, USA. The results will help municipalities to overcome challenges and start well-informed excavation plans. This is expected to reduce the costly unnecessary excavations, minimize disruptions, and promote public health.

2. PROBLEM STATEMENT AND OBJECTIVE

To comply with regulations and protect public health, replacing lead service lines is essential. Municipalities must first identify their number and locations for effective planning. Recently, many states in the U.S. declared that they have 16% of their inventory of unknown or suspected lead (Hensley et al., 2021). The incomplete inventories challenge is rooted in discrepancies in information sharing, poor collaboration and coordination, and a dependency on outdated paperwork and records, many of which, if not lost, are inaccurate and incomplete (Hajiseyedjavadi et al., 2020; Hensley et al., 2021). The tap water test, is another way to determine the existence of lead in the SLs. However, conducting the test is confronted with many challenges, such as the resident's private access denial, test settings, and staff requirements besides being time-consuming and costly, making it complex and ineffective. On the other front, starting an excavation is costly and complex; it needs large machinery tools, resources, and causes service and social disturbances. In addition, wrong excavations result in catastrophic social, economic, and environmental consequences (Hajiseyedjavadi et al., 2020; Hensley et al., 2021; Lytle et al., 2021).

3. PREVIOUS WORKS

Since the attention of the high lead exposure has been raised as a public health concern, many researchers have addressed this issue, aiming to identify and localize the exposure to support interventions. To identify high-exposure occurrences and locations, Zartarian et al. (2022) reviewed multi-sourced data, predictive modeling, and local and environmental indices to identify the locations with high risk, aiming to prepare risk reduction strategies and prioritize intervention efforts. (Wheeler et al., 2021) and (Schultz et al., 2017) explored predictive modeling for lead exposure risk based on neighborhood socioeconomic status SES variables across the U.S. and blood test results. In (Wheeler et al., 2021) Bayesian index model outperformed all for elevated blood lead level (EBLL) risk modeling, and in (Schultz et al., 2017), multiple regression model results were assessed against three states measured BLL data with a maintained R^2 of 0.69, 0.28, and 0.20. Besides, Gould (2009) conducted a cost-benefit analysis study on a lead-hazard control plan to support the efforts. Although being addressed as a problem in public health and associated with blood tests (Bellinger, 2008a, 2008b; Brown & Margolis, 2012), the lead in drinking water is linked directly to the existence of Lead pipes. The Lead pipes are components of the supply water networks that are managed by municipalities. (Gurewitsch et al., 2017) addressed the lead water contamination by lead pipes in Pittsburgh. Through a GIS community-assisted study to identify high-risk houses. Geostatistical analyses of age, income, poverty rate, and other attributes contributed to preventing lead exposure from tap water. The lack of data on full inventories makes the machine learning application suitable to reduce uncertainty in decision-making when addressing lead in water" (Hajiseyedjavadi et al., 2020).

Therefore, some researchers applied machine learning classifiers to predict the service lines class. (Abernethy et al., 2018) documented their ongoing work in Flint for detecting lead service lines for an active remediation plan. They trained and tested three machine-learning predictive classifiers (RF, Lasso, XGBoost) on 6,505 of multi-sourced homes data; the XGBoost outperformed all in terms of accuracy with AUCROC equal to 0.94. The best model was incorporated with a second stage that supports decision-making on service pipe replacement. (Hajiseyedjavadi et al., 2020) developed and applied six different models (LR, AAN, KNN, RF, SVM, GBM) to predict tap water concentrations of more than 15 ppb in Pittsburgh depending on mixed data resorces. The best model was GBM, which maintained an AUROC of 71.6%. The study plans to deploy the GBM model to identify houses with lead service lines to minimize the number of wrong excavations. From another perspective, to detect fraudulent water customers in the water utility (Al-Radaideh & Al-Zoubi, 2018), applied intelligent data mining techniques of two classifiers (SVM and KNN) to detect suspicious fraud water customers. The KNN model maintained a higher accuracy (hit rate over 74%), and the model was deployed in a decision tool to help the company predict suspicious customers to be considered for further inspection. (Omar et al., 2023) Applied data mining to predict water mains' failure in the city of Kitchener. Two scenarios were modeled and evaluated using six prediction models. The random forest model outperformed all with an accuracy of 97.3% and an F1-score of 80.4%. The results revealed that by retrofitting a portion of 8% of the existing network, 72% of breaks could be avoided. Based on this discussion, it is well demonstrated that studies utilizing data mining techniques to address lead service line detection are not sufficient, despite their potential and the existence of big data in open city portals. Besides, various contexts have their own data availability, factor considerations, and circumstances. This has opened the opportunities to conduct more studies that apply different machine learning models and investigate their potential to address this issue.

Given this context, this study applies an innovative methodology to determine lead service lines that need replacement in the city of Aurora. To fulfill this aim, the study will identify the dataset's key attributes that are associated with lead existence and then find the best machine-learning model that can predict the lead service lines. Finally, the study will develop an intuitive tool to be used by the municipality for detecting the LSLs for well-informed replacement plans.

4. METHODOLOGY

This section presents the methodology followed in this study which is based on the CRISP-DM. As illustrated in the main stages of the methodology in Figure 1, it commences with a prioritization process utilizing DBSCAN followed by prediction models. Several models were tested, drawing on those previously developed in the literature, including Naïve Bayes. Among them, three models demonstrated strong performance and were selected for modeling in this study, namely K-Nearest Neighbors (KNN), Decision Tree (DT), and Square Vector Machine (SVM). The following sections will provide each stage in detail.

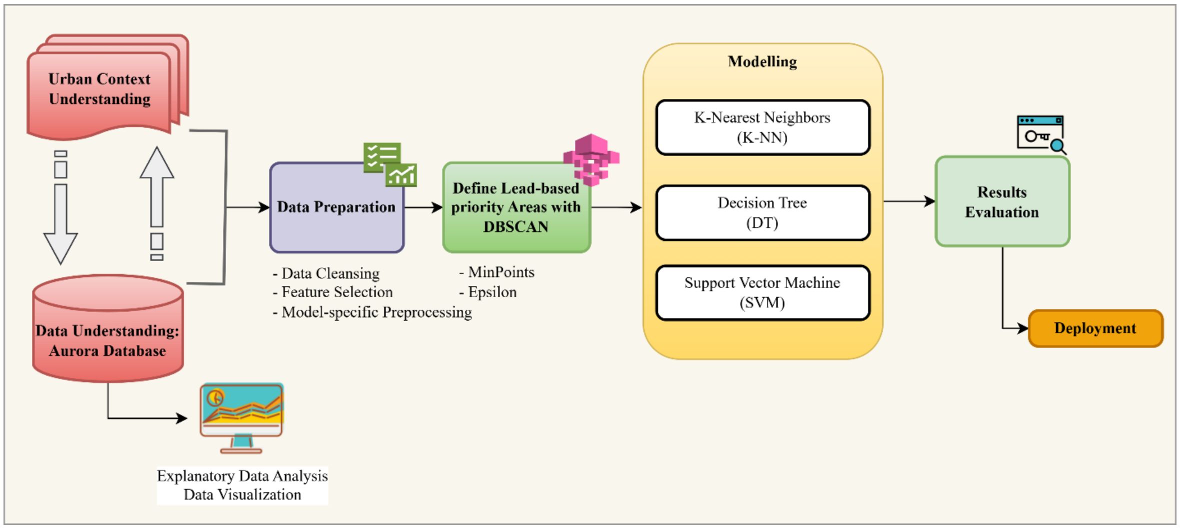

Urban Context Understanding
Data Understanding: Aurora Database
Explanatory Data Analysis
Data Visualization
Data Preparation
- Data Cleansing
- Feature Selection
- Model-specific Preprocessing
Define Lead-based priority Areas with DBSCAN
- MinPoints
- Epsilon
Modelling
K-Nearest Neighbors (K-NN)
Decision Tree (DT)
Support Vector Machine (SVM)
Results Evaluation
Deployment

Figure 1: The schematic flowchart of the proposed methodology

4.1 Understanding Data

The dataset used in this study was obtained from historical and maintenance records of water service material inventory of Aurora city open data portal. The dataset covers the period from 2018 to 2023, a total area of around 150 km^2 and includes information on house location and related water SL materials such as longitude, latitude, service address, house SL material, street SL material and installation year. The obtained dataset comprises 20 attributes (16 nominal, 2 numerical and 2 datetime) as well as one Boolean target variable (Is Lead) and 49632 datapoints. After thorough analysis of the literature to identify the most influential attributes considered in previous studies similar to this work, the irrelevant attributes , such as the date the customer got notified of lead and if the customer refused property access, were removed and the remaining set of 13 attributes and 1 target variable which will be considered in this study are presented in Table 1 accompanied with their data type and description.

Table 1: Attributes and Attribute types

Attribute	Description	Data Type
X	Latitude	Numerical
Y	Longitude	Numerical
ServiceAddress	Address of the parcel/property	Categorical
YearHouseSideServiceLineWasInst	Year of installation of House Side Service Line (SL)	Date/Time
StreetSideServiceLineMaterial	Street-side SL material	Categorical
HouseSideServiceLineMaterial	House-side SL material	Categorical
GalvanizedEverDownstreaofLead	Was Galvanized Service Line Material downstream of Lead?	Boolean
LeadServiceReplacedAtThisLoc	Was a Lead Service Line Previously Replaced at this Location?	Boolean
MaterialOfServiceReplacement	Material of House Side SL Replacement	Categorical
SuspectedLead	Suspected Lead	Boolean
HighRiskFacilityOrArea	Critical Site	Boolean
Gooseneck_Pigtail	Material of Gooseneck/Pigtail	Categorical
SuspectedCopper	Suspected Copper	Boolean
Is Lead (Target Variable)	**SL Contains Lead (Yes/No)**	**Boolean**

Descriptive statistics revealed missing values in SuspectedCopper, Gooseneck_Pigtail, and MaterialOfServiceReplacement, as well as inconsistencies in house SL installation year ("U"). Bar charts and histograms were used to visualize data distributions, identify inconsistencies, and detect outliers. Some attributes, like house SL installation year, were modified for clarity (e.g., converted to "Age"). These visualizations, combined with statistical insights, provided a clearer understanding of the dataset before model training. Figure 2 shows that copper and lead are the most dominant materials used by the municipality for House Side pipes. In addition, street SL material and house SL material attributes have inconsistencies (“unknown”). On the other hand, the Age histogram in Figure 3 shows that although it ranges from 0 to more than 150 years, more than 99% of the data are less than or equal 125 years with an average of 50 years. Therefore, the records which have installation year of house SL more than 125 years were considered as outliers. Finally, more than 63% of datapoints do not have lead which indicates that the dataset is biased towards the “No” class. Therefore, inconsistencies, outliers and missing data need to be prepared and preprocessed before modeling.

4.2 Data Preparation

In this project, the data preparation and preprocessing stage consisted of three key phases. These included data cleansing and scrubbing to enhance accuracy and consistency, correlation analysis and feature selection to understand attribute relationships and identify significant predictors, and ultimately, model-specific data preparation to adapt the dataset according to the format supported by each algorithm. At the outset of the data cleansing and scrubbing phase, the KNN imputer was employed to replace missing values in the "Age" attribute with the mean value from its K nearest neighbors. The Imputer was used to estimate missing values in 'YearHouseSideServiceLineWasInst' by leveraging similarities among neighboring data points. Categorical attributes were one-hot encoded, and numerical attributes were normalized using MinMaxScaler() to ensure consistency. The KNN imputer (n_neighbors=5) then replaced missing values based on the average of the five nearest neighbors. Finally, an inverse transformation restored the original scale, ensuring accurate and interpretable imputed values. Subsequently, all records that contain missing values (e.g., Gooseneck pigtail and galvanized downstream of Lead have missing values for some houses), inconsistencies (e.g., house and street SL material attributes have “unknown” values for some examples), and outliers (e.g., Age has outliers for some records) were filtered out. It is also worth noting that service lines installed after 1990s were carefully reviewed. These records typically featured materials other than lead due to regulatory changes. Therefore, these records served as an internal validation of the model’s ability to differentiate based on regulatory compliance timelines. In addition, any value less than (Q1 – 1.5 IQR) or greater than (Q3 + 1.5 IQR) was considered an outlier, where Q1 and Q3 are the 25th and 75th percentiles, respectively, and IQR is the interquartile range.

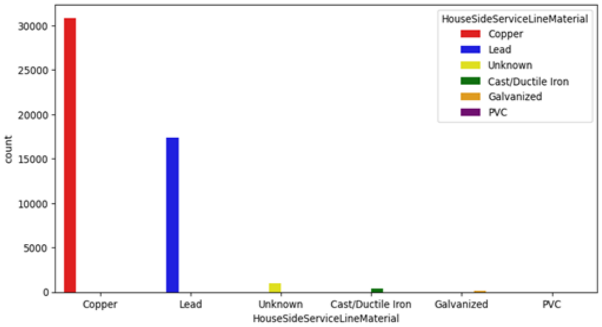


Figure 2: Bar chart- HouseSideServiceLineMaterial SL types and frequency

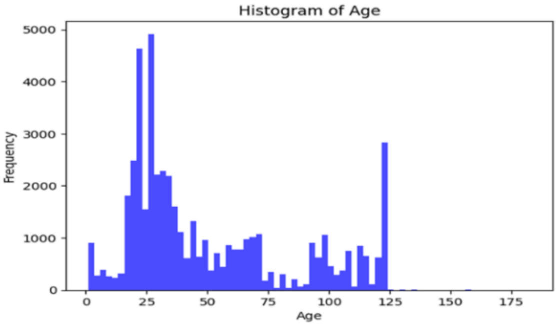


Figure 3: Histogram for Age attribute

The next step involved identifying the optimal set of features to construct an optimized model by highlighting independent variables that are correlated with one another (redundant features). This helps determine the importance level of various attributes in predicting the target variable. Therefore, a correlation matrix was initially generated for both numeric and categorical variables, as depicted in Figure 4. This matrix calculates the strength of correlation between features using Pearson's R (Pearson correlation coefficient) for continuous-continuous cases, Correlation Ratio[1] (Pearson, 1923) for categorical-continuous variables, and Cramer's V[2] (Ester et al., 1996) for categorical-categorical attributes. As depicted in this Figure, the correlation measure among five attributes i.e. "Street Side SL Material," "House Side SL Material," "Gooseneck Pigtail," "Suspected Lead," and "Suspected Copper” falls within the range of 0.97 to 1, indicating a robust correlation among these variables. To elaborate, all these attributes pertain to the physical characteristics (material) of distinct sections within an individual service line. Given the imperative for consistency and uniformity throughout the entire system, the calculated correlation can be rationalized.

Consequently, these five attributes were consolidated and replaced with a singular attribute named "Material Summary”. The MaterialSummary attribute was derived by consolidating multiple material-related attributes, including StreetSideServiceLineMaterial, HouseSideServiceLineMaterial, Gooseneck_Pigtail, SuspectedLead, and SuspectedCopper. To ensure consistency, 'Yes' values in SuspectedCopper and SuspectedLead were replaced with 'Copper' and 'Lead', respectively. The mode function was then applied to identify the most frequently occurring material type across these attributes, ensuring that MaterialSummary accurately represents the predominant material for each record. Having addressed the interdependency among variables, the importance levels of the remaining attributes were determined using the chi-squared test and Decision Tree algorithm. Following the removal of outliers for the "Age" and "Y" attributes and converting numeric variables such as "X," "Y," and "Age" to categorical format through a Bin Discretizing approach, the chi-squared test was applied. According to the obtained results depicted in

[1] Correlation ratio is a measure of the curvilinear relationship between the statistical dispersion within individual categories and the dispersion across the whole population. It varies within the range of 0 to +1.

[2] Cramér's V or Cramér's phi (denoted as φ_c) is a measure of association between two nominal variables, giving a value between 0 and +1 (inclusive). It is based on Pearson's chi-squared statistic.

Figure 5, the two attributes of "MaterialOfServiceReplacement" and “LeadServiceReplacedAtThisLoc” were identified to have the lowest impact (chi score) on the prediction process.

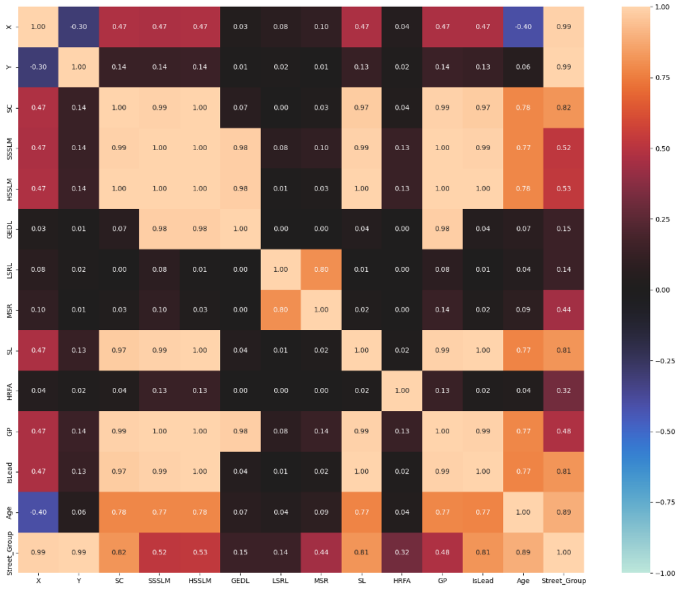


Figure 4: Correlation matrix

(Abbreviations: GEDL: Galvanized Ever Downstream of Lead; LSRL: Lead Service Replaced at This Location; HRFA: High-Risk Facility or Area; MSR: Material of Service Replacement; SC: Suspected Copper; SSSLM: Street Side Service Line Material; GP: Gooseneck/Pigtail; HSSLM: House Side Service Line Material; SL – Suspected Lead)

In addition, the weights generated by the DT, calculated based on the gain ratio criterion, were employed to pinpoint other attributes with limited significance. Illustrated in Figure 6, the attribute "HighRiskFacilityOrArea" was eliminated by the DT, signifying its marginal importance. To conclude, given that the DT ranked the variable "MaterialOfServiceReplacement" as the fourth influential attribute, the other two predictors ("LeadServiceReplacedAtThisLoc" and "HighRiskFacilityOrArea") were excluded from the dataset.

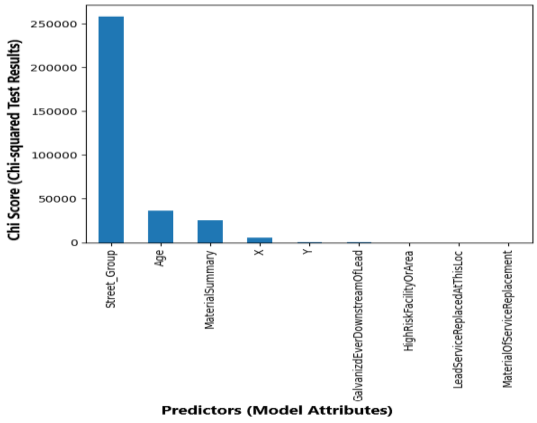
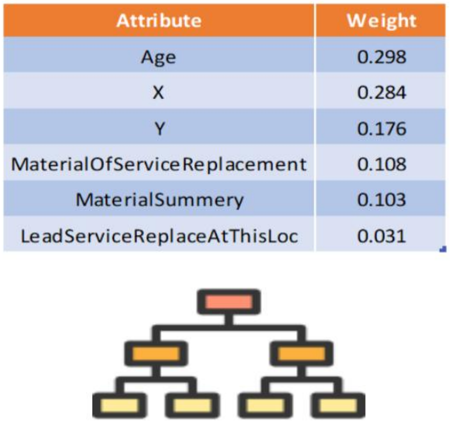


Figure 5: Chi^2 Scores



Attribute	Weight
Age	0.298
X	0.284
Y	0.176
MaterialOfServiceReplacement	0.108
MaterialSummery	0.103
LeadServiceReplaceAtThisLoc	0.031

Figure 6: Decision Tree Attributes Weights

During the model preparation phase, data standardization was applied to rescale the "X," "Y," and "IsLead" attributes before clustering with DBSCAN using range normalization, as the data did not follow a normal distribution. For KNN and SVM, which require numeric and normalized data, one-hot encoding was used for categorical attributes to prevent ordinal relationships. This was particularly important for KNN, a distance-based algorithm, and SVM, which maps data into higher-dimensional spaces for better decision boundaries. Range normalization (0 to 1 scaling) was used for "X," "Y," and "Age" since they were not normally distributed. For Decision Tree (DT), the dataset was split into predictors and the target variable (IsLead). While DT can handle numerical and categorical data, the Python DT module does not support categorical predictors, requiring dummy coding for variables like Material Summary, Material Service Replacement, and Street Group. However, since most attributes were nominal, interpretation of predictor importance was limited. To address this, RapidMiner was used for DT modeling, yielding more reliable results.

4.3 Modeling

To achieve the second objective, DBSCAN was used to identify lead-priority areas, followed by classification models for LSL prediction. This section outlines the models and their hyperparameters. Current project addresses the challenge of managing numerous lead-related complaints in municipalities by developing a prioritization tool. This tool aims to identify high-risk areas efficiently, allowing for prioritization of lead-related complaints before employing more complex classification models, optimizing time and resources. In pursuit of this objective, the DBSCAN algorithm was employed (Birant & Kut, 2007), renowned for its robust clustering capabilities that can discern clusters of different shapes and handle noise effectively. It is noteworthy to mention that the algorithm exhibits limitations in detecting clusters with varying densities or incorporating both spatial and non-spatial attributes (Ashari & Tjoa, 2013). DBSCAN relies on two primary hyperparameters: ε (epsilon) and MinPoints, which regulate the sensitivity of the clustering process and require careful tuning. In this project, the sensitivity analysis of DBSCAN was conducted using the measure of WCSS (Within-Cluster Sum of Square) and a k-distribution chart (based on the distance to the kth nearest neighbor). The optimal values for ε and MinPoints were determined to be 6 and 0.03, respectively, employing the elbow method to identify a point indicating a significant change in WCSS or KNN distance.

Optimized hyperparameters were applied to DBSCAN on the Aurora dataset, revealing six clusters. Clusters 1, 2, 3, and 5 exhibited a positive status for the target variable ("IsLead"), guiding the delineation of lead-based priority zones based on Figure 7. While DBSCAN identified key lead-prone clusters, Priority Area 1 was notably large, which may limit its usefulness for targeted action. This reflects a known limitation of DBSCAN in handling variable-density areas. Future work may explore alternative clustering methods for more refined prioritization. Once the data is prepared, training different algorithms (KNN, SVM and DT) and applying them is the next step. Three different scenarios were modeled in this study: The first scenario includes the most significant attributes after feature selection, with all material related attributes merged as discussed in the data preparation subsection. The second scenario consists of the same attributes as the first scenario. However, oversampling was utilized to balance the training set. The final scenario was modeled to assess how the models perform when the municipality lacks access to complete data, particularly regarding the material of the pipes. Hence, the last scenario (Scenario 3) includes property/parcel related attributes ‘X’ ‘Y’ ‘Age’ ‘Street Group’ only as predictors for the target variable. Initially, the KNN algorithm stands out as a highly efficient non-parametric classification method, characterized as a lazy, distance-based algorithm.

In the KNN technique, predictions are made by examining the entire dataset to identify the K closest (or most similar) records, hence the effectiveness of this model is closely tied to the choice of the K value. By selecting Euclidean distance as a proximity measure and upon fine-tuning the model with different values of K, the optimal number of k was identified as 10. The previous study iterated the values of K from 1 to 10 and achieved the best accuracy when K value was 8 (Al-Radaideh & Al-Zoubi, 2018). SVM is a powerful and versatile algorithm that excels in handling both linear and non-linear classification problems by finding an optimal hyperplane that separates different classes (Cortes & Vapnik, 1995). SVM was tested for all scenarios, providing consistent performance in each case. Conversely, the decision tree algorithm excels in terms of interpretation, visualization, and managing non-linear relationships between predictors and output results, demonstrating excellent performance in these aspects (Hastie et al., 2001). Nonetheless, the challenge of overfitting remains a persistent concern with this algorithm. In order to mitigate overfitting in the model, the gain ratio was employed as a criterion to assess how the impurity of a split would be measured. Through fine-tuning the model with various values, the optimal number of maximum depths was identified as 10.

4.4 Evaluation

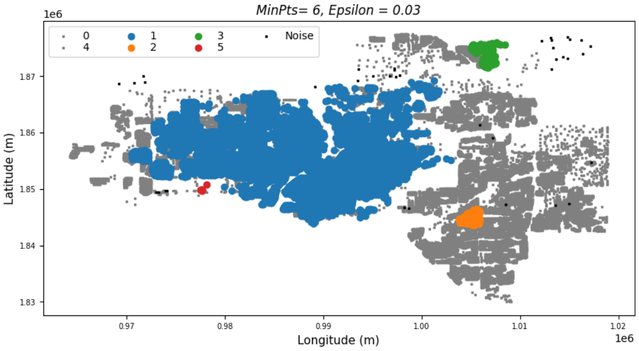


Figure 7: Lead-based priority areas resulting from DBSCAN clustering process

To assess the performance of each model in the various scenarios, accuracy, precision, recall, and F-1 score were evaluated. Additionally, the Receiver Operating Characteristic Curve and the Area Under the ROC Curve (AUC) were analyzed to gauge the goodness of the models. Split validation (80% training set and 20% test set) as well as a 10-fold stratified cross-validation were utilized to assess the models. In this study, false positive rates and false negative rates are both important to investigate, with false positives being particularly critical. Falsely identifying a pipe as containing lead leads to unnecessary excavation and costs, making it a significant concern. Conversely, false negatives, while also problematic, may result in failing to identify a pipe that genuinely requires replacement, posing health hazards to household occupants. Given that false positives have a greater impact, precision is a key factor in model evaluation. However, to ensure a balanced assessment that considers both precision and recall, the F1-score was used to evaluate and compare the models, as it provides a more comprehensive measure of overall performance. In situations involving imbalanced datasets, relying solely on the accuracy metric will potentially lead to misclassification, as accuracy treats each class equally. Hence, a more nuanced and insightful evaluation of model performance can be given by precision, recall, F-1 score. F-1 score in particular illustrates a good balance between true and false predictions in situations where one class is substantially underrepresented compared to the other, which is the case in scenarios 1 and 3 of this study.

5. RESULTS AND DISCUSSION

Table 2 presents the performance metrics for the three models across the three modeled scenarios. SVM excelled in Scenario 1 with an F1-score of 100%, but its performance dropped significantly to 57% in Scenario 2 due to oversampling. DT benefited from oversampling, improving from 94.1% in Scenario 1 to 99.8% in Scenario 2, while KNN remained highly consistent, maintaining F1-scores of 99.0% in both scenarios and showing the highest precision in two out of three cases (99.0%). In Scenario 3, all models declined, with KNN performing best (92.0%), followed by SVM (90.0%) and DT (88.6%). These results highlight KNN's stability, DT's adaptability to oversampling, and SVM's sensitivity to data balancing techniques.

Table 2: Summary of results obtained from applied models

	Algorithm	Accuracy	Precision	Recall	F1-Score	AUC
1st Scenario (Attributes after feature selection)	DT	0.957	0.921	0.962	0.941	0.999
	KNN	0.994	0.990	0.990	0.990	1.000
	SVM	0.999	1.000	1.000	1.000	1.000
2nd Scenario (Scenario 1 + oversampling)	DT	0.998	0.995	0.988	0.998	0.998
	KNN	0.989	0.990	0.990	0.990	1.000
	SVM	0.567	0.610	0.610	0.570	0.730
3rd Scenario (Property/parcel attributes only)	DT	0.917	0.863	0.9129	0.886	0.926
	KNN	0.924	0.920	0.920	0.920	0.970
	SVM	0.903	0.890	0.910	0.900	0.950

It is worth pointing out that the three models, KNN, SVM and DT exhibited F-1 scores of 92.0%, 90.0% and 88.6% respectively in the results of scenario three, where only property/parcel related attributes (X and Y coordinates, age and street group) were used as predictors in the absence of any pipe material related data. This indicates that the prediction model can still be useful in predicting lead presence in cases when the municipalities do not have complete data related to the property, parcel or location at hand. This demonstrates the model's robustness and practicality, particularly for municipalities lacking complete pipe inventories, offering a reliable approach for targeted planning and cost-effective intervention. The results of this study were also compared to similar studies from the literature for each model. For KNN, the results were compared to (Dritsas & Trigka, 2023) which used predictive models to identify the suitability of water for various uses. The comparison is summarized in Table 3 below. The developed models show higher values in all performance metrics especially accuracy and precision where the difference in values between the models and literature is larger.

Table 3: Comparison of goodness metrics with the literature

Authors	Accuracy	Precision	Recall
This study (Scenario 1)	0.994	0.990	0.990
This study (Scenario 2)	0.989	0.990	0.990
This study (Scenario 3)	0.924	0.920	0.920
Dritsas, E., & Trigka, M.[21]	0.886	0.822	0.978

In the practical context, to facilitate the application of the presented methodology for Aurora's municipality, a feedback user interface was introduced (Gradio User Interface Link). The developed tool is an intuitive and user-friendly decision support application tailored for end users (municipal personnel), utilizing the predictive power of the KNN classification model. The interface collects essential information, including longitude, latitude, age, address, and the composition of service line material to predict the existence of lead service lines. It aims at minimizing unnecessary excavations, which are both costly and disruptive. This practical output directly translates to reduced labor costs, fewer service interruptions, and improved compliance with public health regulations.

6. CONCLUSION

The water distribution system is essential for delivering safe drinking water, requiring municipalities to replace lead service lines. This study developed a lead-based prioritization tool using DBSCAN clustering and applied DT, KNN, and SVM models under three scenarios to predict lead service lines. Key predictive features included pipe age, location, replacement material, and material summary. SVM performed best in Scenario 1 (F1-score: 100%), while KNN showed consistent performance across scenarios (99% in Scenario 1, 92% in Scenario 3). Data balancing improved DT slightly but significantly reduced SVM's accuracy, while KNN remained stable. Based on these results, KNN was integrated into a decision-making tool, ensuring accurate lead pipe predictions with minimal misclassifications. This study advances the field by integrating spatial clustering (DBSCAN) with machine learning to predict infrastructure risk, offering a scalable approach for identifying lead service lines, even with limited data. Scenario 3 showed lower accuracy than Scenario 1 but remains useful where data is limited. It demonstrates that location and property-related features alone can yield usable predictions. The model effectively minimizes false positives and negatives, reducing excavation costs and protecting public health. Future work should explore incorporating additional predictive factors such as property characteristics, demographics, and socio-economic indicators to enhance accuracy. Expanding the dataset to multiple cities would improve generalizability and adaptability to different urban environments. To support scalability, the proposed methodology can be adapted to other municipalities that face similar issues of incomplete pipe inventories by incorporating local datasets and retraining the algorithms on context-specific features such as regional material preferences or zoning patterns. Additionally, testing deep learning approaches or ensemble models could further refine predictions. Finally, developing a user-friendly GIS-based decision support system would help municipalities streamline lead service line replacement efforts.

REFERENCES

Abernethy, J., Chojnacki, A., Farahi, A., Schwartz, E., & Webb, J. (2018). *ActiveRemediation: The Search for Lead Pipes in Flint, Michigan*. *4*. https://doi.org/10.1145/3219819

Al-Radaideh, Q. A., & Al-Zoubi, M. M. (2018). A data mining based model for detection of fraudulent behaviour in water consumption. *2018 9th International Conference on Information and Communication Systems (ICICS)*, 48–54.

Ashari, A., & Tjoa, A. M. (2013). Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 4, Issue 11). www.ijacsa.thesai.org

Bellinger, D. C. (2008a). Neurological and Behavioral Consequences of Childhood Lead Exposure. *PLoS Medicine*, *5*(5). https://doi.org/10.1371/journal

Bellinger, D. C. (2008b). Very low lead exposures and children's neurodevelopment. *Curr Opin Pediatr*, *20*(2).

Berglund, E. Z., Shafiee, M. E., Xing, L., & Wen, J. (2023). Digital Twins for Water Distribution Systems. *Journal of Water Resources Planning and Management*, *149*(3). https://doi.org/10.1061/jwrmd5.wreng-5786

Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data and Knowledge Engineering*, *60*(1), 208–221. https://doi.org/10.1016/j.datak.2006.01.013

Brown, M. J., & Margolis, S. (2012). Lead in drinking water and human blood lead levels in the United States. *MMWR Suppl.*, *61*(4).

City of Aurora Open Data. (n.d.). *2023 IEPA Water Service Inventory Submission*.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Leaming*, *20*, 273–297.

Dritsas, E., & Trigka, M. (2023). Efficient Data-Driven Machine Learning Models for Water Quality Prediction. *Computation*, *11*(2). https://doi.org/10.3390/computation11020016

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. www.aaai.org

Gould, E. (2009). Childhood lead poisoning: Conservative estimates of the social and economic benefits of lead hazard control. *Environmental Health Perspectives*, *117*(7), 1162–1167. https://doi.org/10.1289/ehp.0800408

Gradio User Interface Link: . (n.d.).

Gurewitsch, R., Karimi, H., & Naccarati-Chapkis, M. (2017). *Mapping Pittsburgh's Lead Problem*. University of Pittsburgh.

Hajiseyedjavadi, S., Blackhurst, M., & Karimi, H. A. (2020). *A Machine Learning Approach to Identity Houses with High Lead Tap Water Concentrations*. www.aaai.org

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning* (Vol. 1).

Hensley, K., Bosscher, V., Triantafyllidou, S., & Lytle, D. A. (2021). Lead service line identification: A review of strategies and approaches. In *AWWA Water Science* (Vol. 3, Issue 3). John Wiley and Sons Inc. https://doi.org/10.1002/aws2.1226

Lytle, D. A., Formal, C., Cahalan, K., Muhlen, C., & Triantafyllidou, S. (2021). The impact of sampling approach and daily water usage on lead levels measured at the tap. *Water Research*, *197*. https://doi.org/10.1016/j.watres.2021.117071

Omar, A., Delnaz, A., & Nik-Bakht, M. (2023). Comparative analysis of machine learning techniques for predicting water main failures in the City of Kitchener. *Journal of Infrastructure Intelligence and Resilience*, *2*(3). https://doi.org/10.1016/j.iintel.2023.100044

Pauli, B. J. (2020). The Flint water crisis. *Wiley Interdisciplinary Reviews: Water*, *7*(3). https://doi.org/10.1002/WAT2.1420

Pearson, K. (1923). *On the Correction Necessary for the Correlation Ratio η* (Vol. 14, Issue 4). https://about.jstor.org/terms

Schultz, B. D., Morara, M., Buxton, B. E., & Weintraub, M. (2017). Predicting Blood-Lead Levels Among U.S. Children at the Census Tract Level. *Environmental Justice*, *10*(5).

Wheeler, D. C., Boyle, J., Raman, S., & Nelson, E. J. (2021). Modeling elevated blood lead level risk across the United States. *Science of the Total Environment*, *769*. https://doi.org/10.1016/j.scitotenv.2021.145237

Zartarian, V., Poulakos, A., Garrison, V. H., Spalt, N., Tornero-Velez, R., Xue, J., Egan, K., & Courtney, J. (2022). Lead Data Mapping to Prioritize US Locations for Whole-of-Government Exposure Prevention Efforts: State of the Science, Federal Collaborations, and Remaining Challenges. *American Journal of Public Health*, *112*, S658–S669. https://doi.org/10.2105/AJPH.2022.307051