Arab American University- Jenin

Engineering and Information Faculty

Prediction of Pipes Break in Water Distribution System Using Data Mining Tools

"Case Study Nablus Municipality"

By

May Khalid Shawkat Saleh

Supervisor

Dr. Amjad Rattroot

This thesis was submitted in partial fulfillment of the requirements for the Master's degree in computer science

January/2018

Prediction of Pipes Break in Water Distribution System Using Data Mining Tools

"Case Study Nablus Municipality"

By

May Khalid Shawkat Saleh

This thesis was defended successfully on 20/1/2018 and approved by:

Committee members Signature

1. Supervisor Name: Dr. Amjad Rattroot …………………………….
2. Internal Examiner Name: Dr. Mohammad Awad …………………….
3. External Examiner Name: Dr. Subhi Samhan ………………………

Declaration

This is to declare that the thesis entitled " Prediction of Pipes Break in Water Distribution System Using Data Mining Tools Case Study Nablus Municipality” under the Supervision Dr. Amjad Rattroot is my own work and does not contain any unacknowledged work or material previously published or written by another person, except where due reference is made in the text of the document.

Dedication

To our greatest and most honoured prophet Mohamed - May peace and grace from Allah be upon him-

To my beloved Mother and Father who have always graced me with love and support every step of the way.

To my beloved son, brothers, sisters and friends I dedicate this thesis.

Acknowledgments

First and foremost, all praise is due to Allah, the Most Merciful, the Most Compassionate, and the author of knowledge and wisdom, for the help and foresight that enabled me to complete this work.

I also extend my supervisor, Dr. Amjad Rattroot my sincerest sense of gratitude for his constant encouragement and invaluable insights. His constructive comments and suggestions contributed to the completion of this thesis. The completion of this thesis would not have been possible without his excellent guidance.

Special thanks are also due to the staff members of the Computer Science Master Program at the Arab American University for their efforts.

I also sincerely acknowledge the support of Dr. Subhi Samhan and the staff of the Palestinian Water Authority and MEDRC Water Research who were fully supportive throughout the period of writing this thesis.

ABSTRACT

The problem of water loss from water distribution networks is a major economic problem that worries many stakeholders in the Middle East area, especially those working in municipalities responsible for water distribution networks. Since there is a shortage of water in most countries of the world, this problem occupies great focus in the world especially in major cities, where the socio-economic cost of water loss is increasing.

The breaking of pipes in water distribution networks is one of the main reasons for the loss of water from the network, so there is an urgent need to control this problem to prevent water leakage from the pipes by continuous repairing and maintaining the pipes before the break. Therefore, there is a need to analyze and understand the data related to water distribution networks and to use this data in predicting the breaking of the pipelines and identifying the factors and variables that lead to break before broken pipes.

using of classical mathematical and statistical tools in identifying the parameters which play a major role in the prediction of pipes' break patterns is a complex task; because of the complexity of this system so that this research seeks to create an alternative model that is to be used for predicting pipes' breaks in water distribution networks and for identifying the variables that cause such breaks.

In this research, the applied dataset collected from the water distribution system in the Municipality of Nablus, which is one of the larg cities in the northern West Bank area of Palestine that was taken as a case study. The R language was used to implement seven classification models for pipes break prediction depending on three data mining techniques that are Decision Tree, Logistic Regression and Support Vector Machine.

The first three models were built by using one of these three techniques, then four new models have also been built by combining the two of these techniques.

Comparing the performance of these models shows that the new model that is built by combining the Logistic Regression and Support Vector Machine techniques, which is called LRSVM model that is most reliable model in the anticipation of pipes' breaks because it gave the best values for most of the calculated performance measures as its error rate varied between 0.01 and 0.12, and it may be able to save up to 0.97 water from the amount of water lost from the network, with an accuracy rate that may reach 0.99.

LIST OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CP-complexity

CSV- a comma separated values file

DI- Ductile Iron

Diam- Pipe Diameters

dis_1- New water distribution network

dis_2- old water distribution network

DM- Data Mining

DT- Decision Tree

DTLR- Combination between Decision Tree and Logistic Regression

DTSVM- Combination between Decision Tree and Support Vector Machine

FN- False Negatives

FP- False Positives

Glm- Generalized Linear Models

GPL- General Public License

GS- Galvanized steel

HDPE- High-density polyethylene

KDD- Knowledge Discovery in Data

Ksvm- Kernal Support Vector Machine

LR- Logistic Regression

LRDT- Combination between Logistic Regression and Decision Tree

LRSVM- Combination between Logistic Regression and Support Vector Machine

Nsplit- Number of Splite

pipe ID- The Identify Number for Pipe

PP- Poly Propline

Pressure_z- Pressure Zone

PVC- Poly Vinyl Chloride

Rattle- The R Analytical Tool to Learn Easily

RBF- Radial Basis Function Kernel

Rel Error- Baseline Erorr

ST- Steel

SVM- Support Vector Machine

TN- true negatives

TP- true positives

xerror- cross-validated error

Chapter One

Introduction

1.1 Introduction

Water shortage is a major global problem that is especially concerning in the Arab world, which is experiencing a rapid population growth, with limited freshwater resources and poor water management. In 2011, approximately 75% of the Arab population suffered from water shortage, and nearly half of them lived below the level of extreme water poverty less 100 m/y per person for all population, and this percentage is increasing significantly (UNESCO, 2015).

Before looking into water loss from water distribution networks, we must have a fully-fledged idea of the parts and structures of the networks in order to determine the requirements and data needed to carry out this research Water distribution networks are a wide range of water pipes, which start from the main water distribution tank belonging to any municipality and end with the points of consumption at houses, schools, industrial facilities and others. In general, water distribution networks consist of three types of pipes classified according to their function into: transmission pipes, main distribution pipes and service pipes. (Arasmith, 2009)Transmission pipes are pipes that carry water from the main reservoirs to the distribution networks. The main distribution pipes transport the water from transport pipes and distribute them throughout the city. Service pipes are the pipes that branch out from the main distribution pipes, which transport water from them to the consumption sites. (Arasmith, 2009)

One of the main reasons for the loss of water in water distribution networks is broken water pipes, which cause water leakage until the break is repaired (Jabari, 2017). Therefore, water loss due to broken pipes should be reduced or prevented through the analysis of water distribution network systems and the identification of the factors affecting the water pipes and leading them to break.

As a result of technological development and technology incorporation in all life aspects, it was possible to utilize technology to design models for water pipe management in water distribution networks. Due to the increasing impact of this problem on the world, it has attracted the attention of many researches try to find ways to improve the management of limited water resources reduce their loss and exploit them properly. From this point, the problem of water loss from water distribution networks has been a hot topic for researchers and this thesis is but an extension to the previous research on this area (Rattrout & Saleh, 2016).

The water distribution system is a complex physical system, therefore the processes causing pipe breaks are often difficult to describe using classical mathematical tools, thusly predictive models are usually used instead. These predictive models can be essentially classified into three categories: physical methods, statistical methods and data mining methods. A number of researches conducted studies on the problem of pipe breaking and developed predictive models to try to anticipate pipe breaks before they happen. (Hand, Mannila, & Smyth, 2001) and (Witten & Frank, 2005). (Han & kamber, 2012)

In this research data mining methods, which have achieved great success in the field of predictive models for complex systems, are used to build models that anticipate pipes' breaks.

1.2 Research Problem and Motivation

Palestine is a country with many natural water sources, nevertheless, it suffers from a lack of water supply, due to a number of factors most notably Israel occupation which controls major water sources, fluctuation and lack of rain water in recent years, increasing water losses in the water distribution networks, and the lack of material and economic resources to desalinate seawater in order to extract freshwater.

The problem of water loss from water distribution networks is one of the most serious problems leading to water shortage in Palestine. Some studies showed that the rates of water loss resulting from it are higher than 30% (Jabari, 2017), while other studies argued that in some Palestinian areas this percentage reaches up to 50%. (Ziad Mimi, 2004) , (Adel Al-Salaymeh, 2009) and (Jad Isaac, 2015)

Pipe breakage is one of the main causes of water loss from water distribution networks. Water lost from water distribution networks through leaking pipes in the West Bank in 2013 were as follows: 14% in Ramallah and Al Bireh governorate, 49% in Jerusalem governorate (Jad Isaac, 2015), as shown in Figures (1).

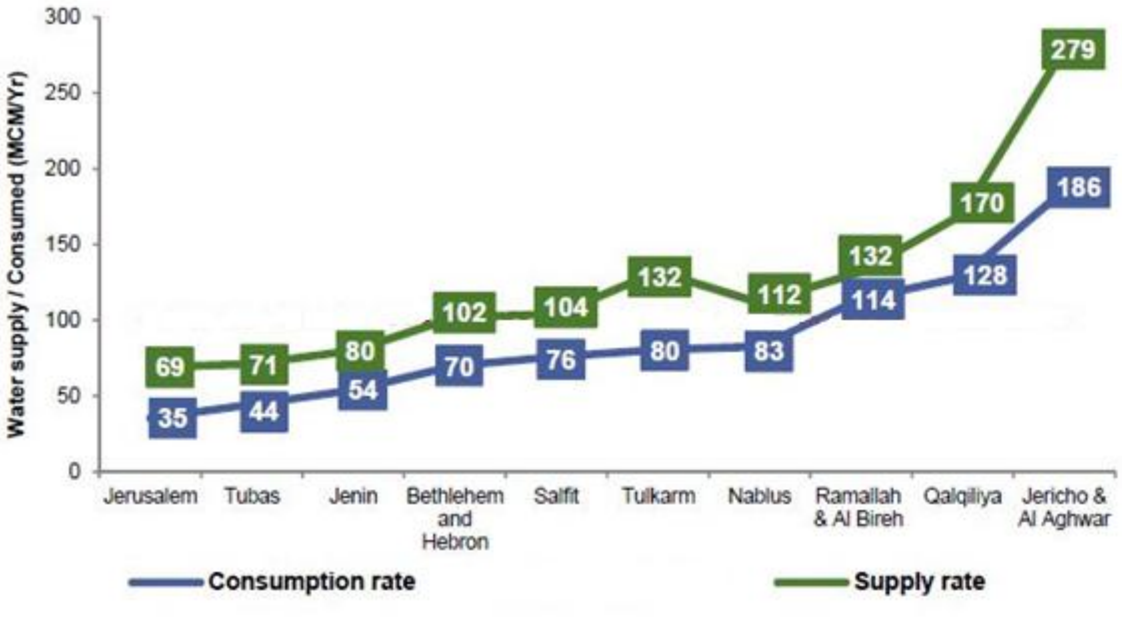
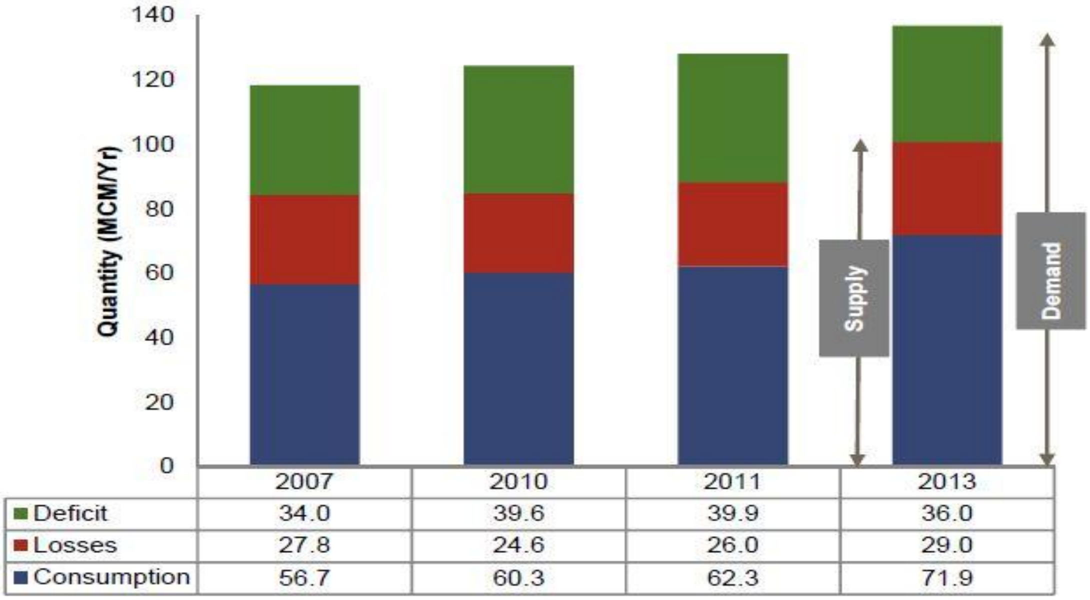


Figure 1: water supply and consumed in west bank year 2013

Figure (2) shows the data related to the consumption, losses and deficits of water in West Bank governorates between the years 2007 and 2013 (Jad Isaac, 2015). Water deficit resulting from water loss comprises somewhere between 65% to 82% of the total water deficit. Therefore water loss should be reduced or prevented from water distribution networks to help resolve the water crisis.



	2007	2010	2011	2013
🟩 Deficit	34.0	39.6	39.9	36.0
🟥 Losses	27.8	24.6	26.0	29.0
🟦 Consumption	56.7	60.3	62.3	71.9

Figure 2: consumption, losses and deficits of water in West Bank governorates between the years 2007 and 2013

Municipalities in all governorates in the West Bank seek to solve this problem by trying to detect and repair broken pipes in the shortest time possible, but in practice this solution is not practical, especially since it takes time to discover the location of the broken pipes and repair them.

The motivation behind this research can be ascribed to the need to reduce water loss, especially loss resulting from broken pipes by solving the problem of broken pipes before it occurs. This is possible when the pipes that are expected to be broken are replaced or repaired prior to their fracture in order to avoid water loss. The study will rely on available data in the Water and Maintenance Departments of the Municipalities in Palestinian governorates.

1.3 Research Question and Objectives

1.3.1 Research Question

When considering the problem of water loss from the water distribution network and trying to solve it, a number of questions arise:

- Can a pipe break be predicted?
- Is it possible to determine the number of pipes expected to break during a certain year?
- What are the factors that cause broken pipes?
- What is the priority and impact of each of these factors?
- Is it possible to predict the amount of water lost from the water distribution network?

1.3.2 Research Objectives

The objectives of this research can be summarized as following:

- Building a new integrated database for pipes in water distribution networks in Nablus, which contains all the data that related to all pipelines in the network as well as maintenance data.
- Finding an effective model to predict pipe breaks in the water distribution network.

- Making a good and effective decisions for pipes maintenance in the water distribution network

Reducing the loss of water from water distribution networks

1.4 The Research Obstacles

Upon commencing this study, the researcher faced some obstacles that surfaced during the data collection phase; those obstacles mostly pertain to a general shortage of data for some pipes, as well as the lack of maintenance records of many years, especially the maintenance data for the years 2012 and 2013.

1.5 Overview of Research Methodology

The data mining techniques are used to achieve the objectives of this research; therefore data mining methodologies is the one to be opted for in this research. The first step of the work after gets a comprehensive grasp of the the problem is data collection. The data is then analyzed, cleaned, integrated, and transformed to an appropriate format so that they are ready for the next step of building the model, where three different classification algorithms are used to build seven prediction models for breaking pipes and to classify the pipes according to their breaking possibility into two classes that is Yes or No.

Then, the performance of the models that were built is evaluated by using appropriate performance measures and the results of the performance are compared and the best model from that is chosen.

1.6 Research Structure

This research is divided into five chapters that can be summarized as following:

Chapter one presents the introduction to the research, which shows the problem of study, its objectives and its importance, and presents an overview of the research methodology.

Chapter two presents a literature review, which has been relied upon as the background for the implementation of this research and which presents some previous research related to the subject of the study.

Chapter three present the details of the research methodology, which were applied in this research, where the knowledge discovery in the data methodology was used in this research to develop three predictive models for pipe breaking by using data mining techniques through the R language.

In Chapter four, the results of predictive models that were constructed are presented, the models performance is measured and the models are compared to reach to the best model.

Chapter five present the summary, recommendations and future works.

Chapter Two

Literature Review

2.1 Introduction

To build any successful and accurate model in any field of life, this model must be well planned to it and relied upon to scientifically proven principles and concepts. This chapter presents the background and all the theoretical concepts necessary to build a precise and effective predictive pipe breaks model. Where it presents all the techniques on which a model building is based and evaluates, and explains the algorithms and methods chosen for this model.

2.2 Background

The background for this research is display in the following sections that are the data mining techniques and algorithms, in addition to models performance metrics.

2.2.1 Data Mining

Nowadays we live in a world are producing massive amounts of data per day, but cannot to benefit from this data without analysis, so that we are need to analyze this data and discovering knowledge from it. With this huge and growing amount of data, there was a need to develop powerful techniques and tools to analyze data and extract information and knowledge thereof. From here came the so-called data mining (Han & kamber, 2012), as a technique designed to analyze data and search on it; to find meaningful patterns of knowledge. Data mining has gained widespread fame and popularity (Kapoor, 2014) and is one of the fastest growing fields in the field of computer science

Data mining can be performed on any form of data to obtain knowledge pattern. Data mining is an essential step in applying intelligent tools on data; to detect important data patterns and hidden in large sets of data, which were not visible before that. Data mining is widely used in diverse areas such as financial, telecommunication, retail, intrusion detection, social media and other scientific applications.

The Data mining "is the computing process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems" (Chakrabarti, et al., 2006).

Some researchers see the term data mining as synonymous with the term knowledge discovery in data (KDD); because it has become more popular to refer to the process in which the whole discovery of knowledge in data. but others believe that the term discovery of knowledge in data (KDD) is more comprehensive and consider data mining a step in the KDD process (Han & kamber, 2012).

Some believe that knowledge discovery in data is an easy process that stops when data is collected and managed, but this is wrong, it extends to analysis and predicting what will happen in the future

The data mining is given as a step in the process of discovering knowledge as illustrated in the Figure (3). The stages and steps of knowledge discovery (Han & kamber, 2012) can be summarized as follows: Data discovery and collection, cleaning, integration, selection, transformation, Model building and validation (data mining), evaluation, and deployment (knowledge presentation).

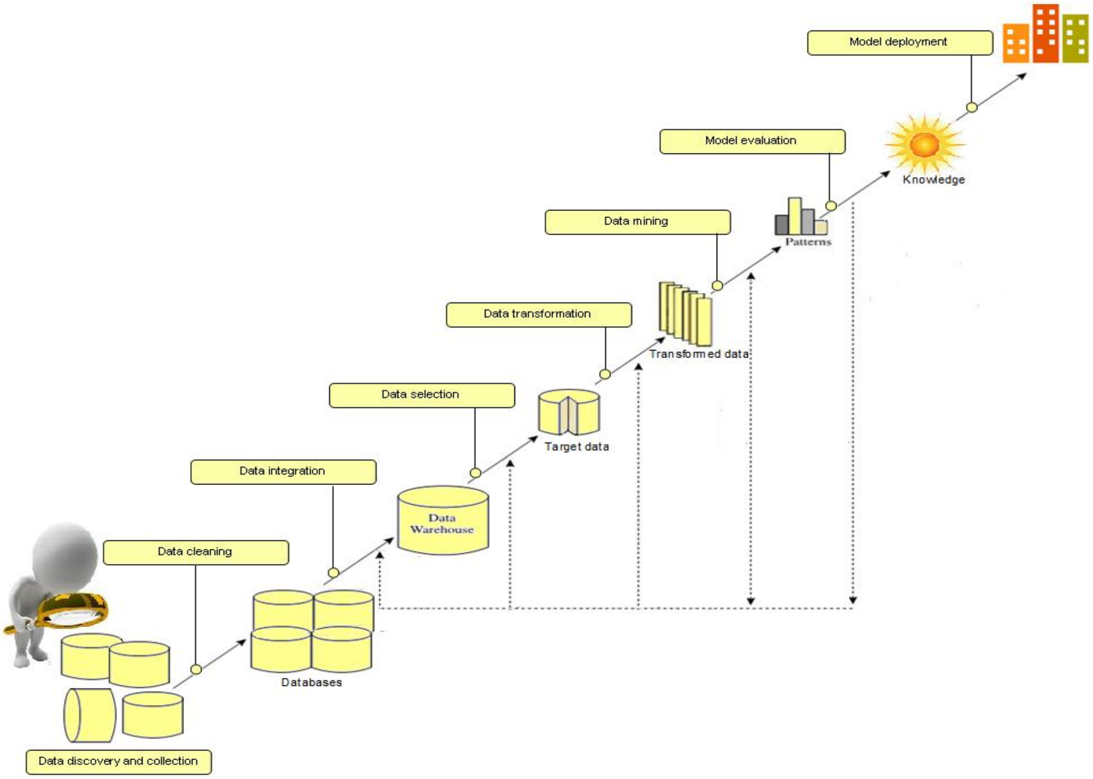


Figure 3: KDD process

Data mining models are divided in terms of the task of data mining into two types (Hamalainen, 2006) (Han & kamber, 2012), (Tamilselvi & Kalaiselvi, 2013), (Sondwale, 2015) , are: Predictive and descriptive models as show in Figure (4).

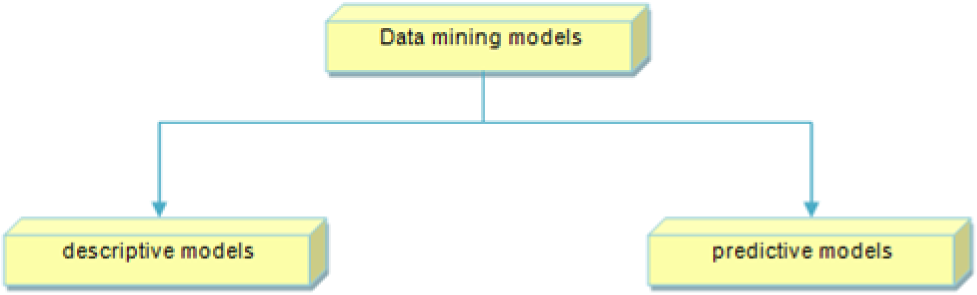


Figure 4: Data Mining models types

Predictive Models are models that are consistent with the method of Elicitation. Where available data are studied and analyzed to predict what the system will be in the next period. That is, old data is used to find the best predictions to predict what will happen in the future of the system (Velickov & Solomatine, 2000).

The predictive model is constructed by forming a model of the target variable with one or more variables in the system. The predictive model is the link between what is known and what is unknown; it predicts the future event as a function of what is now known, which is built using past examples of the future event and is more effective when it relies on a large amount of reliable and correct data.

The appropriate method or technique is chosen to construct the predictive model by determining the data mining task, there are four main methods that can be used in building this type of model (Kalechofsky, September, 2016), are classification, regression, prediction and analysis of time series, as show in Figure (5).

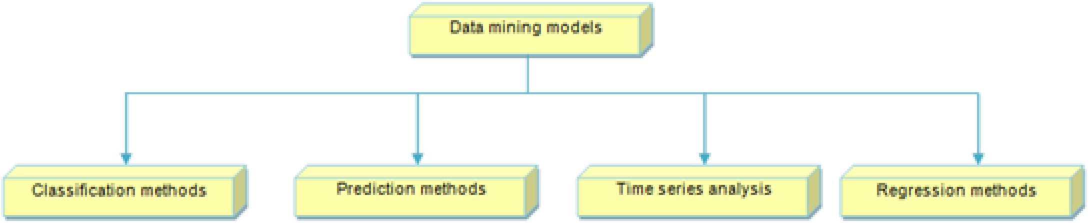


Figure 5: Data mining methods of Predictive Model

Descriptive models are models that are consistent with the method of extrapolation. That is, it uses logic to describe and analyze the relationships between variables affecting the problem, in other words, models that rely on the reorganization of data and exploration of data in depth to extract the indicators in them (Hand, Mannila, & Smyth, 2001).

This type of model is divided into two main categories: clustering, where data and events are grouped into clusters, and correlation models that determine the relationships between variables affecting the problem.

Based on the data mining task of building model the appropriate method is chosen to construct the descriptive data model. The most important methods can be used to build descriptive models are summarizations, clustering, sequence discovery and association rules. As shown in Figure 6.

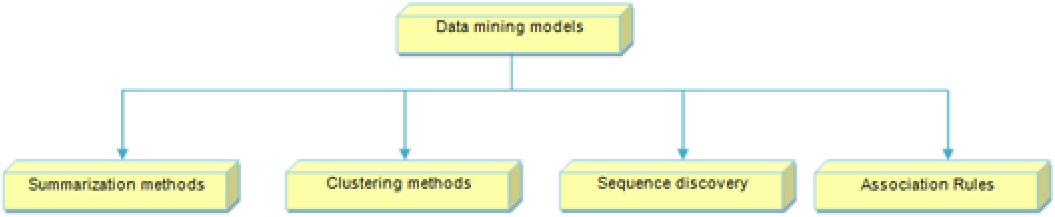


Figure 6: Data Mining Methods of Descriptive Model

2.2.1.1 Selected Model and Method

The predictive data mining model is the appropriate model for use in this study and the classification technique particularly, where classification technique is appropriate, easy and very effective to describe and predict of binary classes (Tan, Steinbach, & Kumar, 2006). The classification technique is one type of supervised learning, which is to interpret or predict the data property through other properties. In the classification technique, a set of classes are predefined, which presenting the object classes of the model, then the classification model is want to find which class the new data belongs to (Tan, Steinbach, & Kumar, 2006)

The classification technique can be applied to the set of input data, to build a predictive model using a number of methods, including: decision tree, support vector machine, naive bayes classifiers, neural network, rule based classifiers (Tan, Steinbach, & Kumar, 2006), random forest (Liaw & Wiener, 2002) and other.

In this research, three algorithms were chosen to build predictive models for pipes break, which are Decision Tree, Support Vector Machine and Logistic Regression algorithms

2.2.1.1.1 Decision Tree (DT) Technique

The decision tree is particularly important in the analysis of decision issues, which contain a series of decisions or successive nature states.

The decision tree is a graphical, Quantitative and graphic representation of the decision-making process, which helps to see all branches of the decision-making process, illustrating all system situations and possible scenarios for decision making (Prajwala, 2015), (Rokach & Maimon, 2014).

The decision tree represents the basic decision as well as secondary decisions that can be subdivided into more specific secondary decisions, depending on specific probability ratios.

That is, they take all possible alternatives to the decision according to the probability of a certain achievement.

That is, they take all possible alternatives to the decision according to the probability of a certain achievement. The basic decision, secondary decisions and related sub-decisions are represented in the form of a tree and its branches.

The decision tree method is the guide or router for the decision-maker towards the statement of that branch of the tree which can lead to the appropriate decision.

The most important feature of a decision tree method to solve problems is that the process of translating and displaying the possibilities of each branch, is done at the end of the tree and Its edges and then are back to its beginning according to the retrograde way.

In other words, the decision to solve a problem begins with the distant goals of the tree, and then the decision-making process moves from a sub-decision to another sub-decision that is closer to the root of the problem, until reaching the final stage, through which everything related to the problem is revealed.

It is worth noting that the decision-maker, relying on this method, is able to understand the problem under study and make decisions about it (Rokach & Maimon, 2014).

Although the method of decision tree is used in the exploration and preparation of data for statistical processes and problem solving, studies and research proved to be one of the best ways to build data mining models (Wu, et al., 2008), which proved its efficiency in classification and prediction techniques

Building the decision tree is not done arbitrarily, but according to the rules and steps specific and clear. The basic rule in building the decision tree is to find the best question (condition) in each branch of the tree, this question divides the data into two parts, the first section of which applies to the question and the second section does not apply to the question. Thus, through a series of questions that represent the conditions of data division, the decision tree is built with its serial branches (Bordley, 2002), (Rokach & Maimon, 2014).

The decision tree consists of a specific set of nodes, where a distinctive node located at the top of the tree is called the root node of the tree, which has no incoming edges. The edges come out from the root node, each edge is returned to another one node. The tree may branch into several levels, where the node can be divided into two more nodes, forming a tree called sub-trees, the node that branch out (from which the edges come out) called the internal node. The node that does not branch out (from which the edges does not come out) called the leaf node (or terminal node), which represent class or decision-making. Figure (7) shows the general shape of the decision tree.

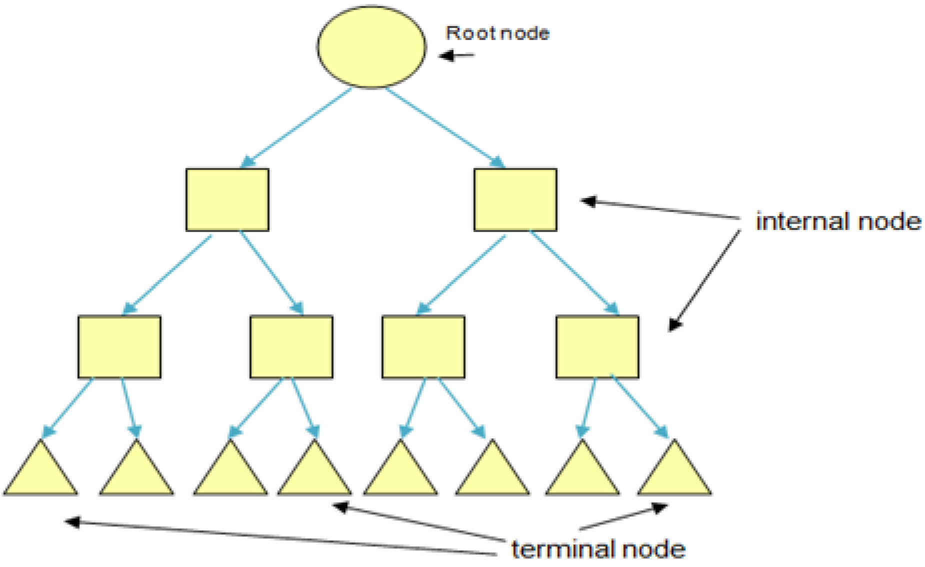


Figure 7: the general shape of the binary decision tree.

When using the decision tree in predictive models, to determine the expected class, start from the root and then visit all the child nodes to which the condition applies on it to reach the target class.

Whenever the data used to construct the model is correct, accurate and reliable, then the decision tree will be fully and correctly expressed about the system that makes it a key factor in reaching the right decision. An information gain measure is used to decide the alternative splits in Decision tree, which use the mathematical logarithm function for base 2 (log2) to transform information. Base 2 is chosen since we use binary classification used.

The formula is used to calculate the entropy of a dataset that is the sum of the probability of each label times the log probability of that same label, is

$$info(D) = -p\ \log_2(p) - n \log_2(n)$$

Where $info(D)$ is entropy, D is training dataset, p is possible value of observations is positive and n is possible value of observations is negative.

Each choice of a split results in a tow partition of the training dataset that is *D1* and *D2* the information measure can be applied to each of these subsets to give I_1 and I_2. A measure of the combined entropy or information can be calculated by the following formula:

$$info(D,S) = \frac{|\,D1\,|}{|\,D2\,|} I_1 + \frac{|\,D1\,|}{|\,D2\,|}\ I_2$$

Where $info(D,S)$ is combined entropy, *D1* and *D2* are a tow partition of the training dataset and I_1 and I_2 are new subset of tree.

The Information Gain that is representing the knowledge, which is the difference between the entropies before and after can be calculated by the following formula:

$$gain(D,S) = info(D) - info(D,S)$$

Where $gain(D,S)$is knowledge, $info(D)$ is entropy and $info(D,S)$ is combined entropy.

2.2.1.1.2 Support Vector Machine (SVM) Technique

The SVM technique was introduced in the early 1990s by the Vapink researcher (Boser, Guyon, & Vapnik, 1992).

SVM is one of a machine learning techniques, specifically is a supervised learning technique, which is based on the statistical learning theory. Support vector machine was created to solve pattern recognition issues, by determining the hyperplane for the data to be separated.

Support vector machine technique is become an important and very effective method in the field of classification (Bhavsar & Panchal, 2012), prediction (Das & Padhy, 2012) and regression in machine learning techniques (Support Vector Machines for Classification and Regression, 1998) and (Wang L. , 2005)

SVM is a useful way to linear data classification, and is also a powerful methodological science for solving problems in nonlinear classification. In the simplest case, the SVM is a binary classification, that is, it separates only two things. There are additions that make SVM capable of dealing with more than two classes; it is also through additions that can deal with regression and multi classification (Mayoraz & Alpaydın, 1999).

The classification process by SVM involves two basic stages, which are the training phase and the testing phase.

Training stage where training data is provided, where each element is inserted have the class and attributes

SVM work to the classification process (Han & kamber, 2012), depending on the training dataset, the SVM model is constructed to predict of class for cases that contain only attributes, which are not known its class.

In the testing phase, the model is implemented on testing data to validate the models. The main goal of SVM is to find the best hyperplane for the data to be separated and categorized.

Hyperplane is the dividing line between the data represented in space, which separates it into two classes in the simplest cases of SVM as shown in Figure 8, and is a surface in the space have dimensions largest than two dimensions.

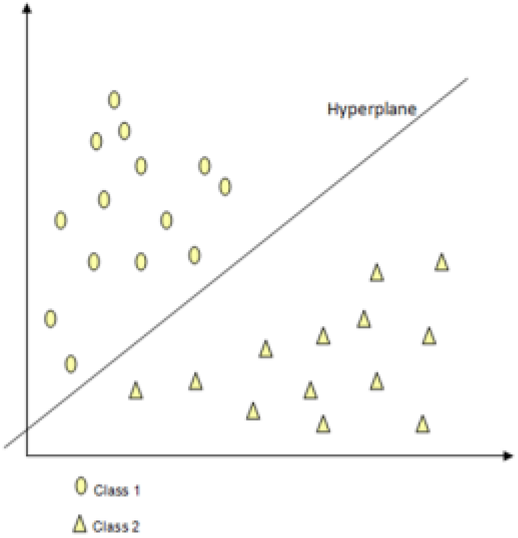


Figure 8: Hyperplane for 2- Dimensions space

The separating hypreplane can be calculated by the following equation:

$$w.x + b = 0$$

Where w is the weight vector of input variables that is $w = \{w_1, w_2, \dots \dots w_n\}$ where n is a number of variables. x is the set of training data with associated class and b is a scalar bias.

To determine the label for each training tuple the following equation is used:

$$y_i = \begin{cases} w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n \geq 1\,, & y_i = +1 \\ w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n \leq 1\,, & y_i = -1 \end{cases}$$

Where y is the class, w is the weight vector of input variables, n is a number of variables and x is the set of training data with associated class

The SVM also known as maximize margin classifiers, where it works to get maximum geometric margin and reduce the empirical classification error.

The maximum margin equation is:

$$MMH = \frac{2}{||w||}$$

Where MMH is maximum margin, $||w||$ is an Euclidean norm of w that is

$$\sqrt{w.w} = \sqrt{w_1^2 + w_2^2 + \cdots + w_n^2}$$

Where w is the weight vector of input variables and n where n is a number of variables.

The maximum margin hyperplane can be achieved by drawing an infinite number of hyperplane The SVM algorithm chooses the hyperplane that is located in the middle exactly in order to get the best results in the classification process. Where the hyperplane that has the largest margin between it and the nearest vector in this space is chosen as shown in Figures 10.

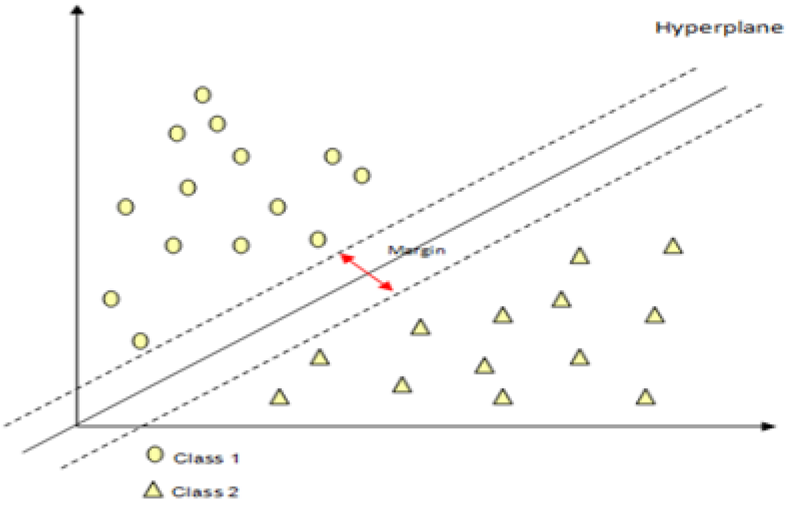


Figure 9: Large margin

In most cases, when data is represented in space, data cannot be separated using straight hyperplane because some data belonging to a given class exists between the data in the other class in space. Therefore, the SVM algorithm has been modified to solve this problem, so that some points are allowed to be placed in the wrong place with no effect on the final result, through the so-called soft margin.

In previous cases, the separate linear data are used, but in some cases the data are represented as vector in space are linear but inseparable. In these cases, the vectors cannot be separated by a straight hyperplane, as the separation between the items is in the form of a curve.

This problem can be resolved by using the kernel functions, which works to transfer data into a multi-dimensional space. The kernel function works to transfer data into a multi-dimensional space, using non-linear representation of data, and then searching for hyperplane to find a maximizing margin using non-linear hyperplane. The kernel functions match between the performance and accuracy of the classification process.

There are many kernel functions that help to represent linear data in a multi-dimensional space are: polynomial kernel of degree n, sigmoid kernenl and gaussian radial basis functions.

polynomial kernel of degree two (Quadrature function) is the simplest type of kernel functions, where the data values are squared and the square values are added as new dimensions in space.

But this function is not used, because it makes the classification process very complicated, because it doubles the dimensions of space, where the complexity increases exponentially as space dimensions increase.

2.2.1.1.3 Logistic Regression (LR) Technique

A logistic regression is a model or statistical method used to predict the probability of a particular event based on the study of a number of independent variables. The binary logistic regression, which is also known as binomial logistic regression, is considered one of the most important models which can be used to study two-valued events, where it is used to examine the relationship between a binary variable and an independent single variable or set of variables of any type. (David G. Kleinbaum, 2010) And (Walsh, 2016),.

The binary logistic regression model assumes that the dependent variable that the model is concerned with it is a binary variable that takes one of the values (1) and (0), where the value (1) represents the occurrence of the event and the (0) value represents the non-occurrence of the event. The probability of the dependent variable is 1 represents by symbol P, the probability of it is 0 represents by symbol (P-1). (Park, 2013)

The logistic regression model can be defined by the following equation: (Park, 2013)

$$p(y) = \frac{1}{(1 + e^{-(p0 + b1\,x1)})}$$

Where $p(y)$ is the probability of Y occurring, e is the natural logarithm base, b_0: is the intersection point with the y-axis, b_1: is the regression coefficient and x_1: is the predictor variable

The regression coefficient determines the effect of the independent variable on the outcome of the dependent variable (the target variable), since the positive regression coefficient means that this variable increases the probability of the dependent variable. If the coefficient is negative, it reduces the probability of the dependent variable. The higher value of coefficient regression, mean the greater effect on the dependent variable (Lorenz, 2015).

Logistic regression is represented in the Cartesian level in the form of logarithmic function, as show in Figure (10), where the x-axis represents the independent variable, while the P(Y) representing on the y-axis, which takes values between 0 and 1.

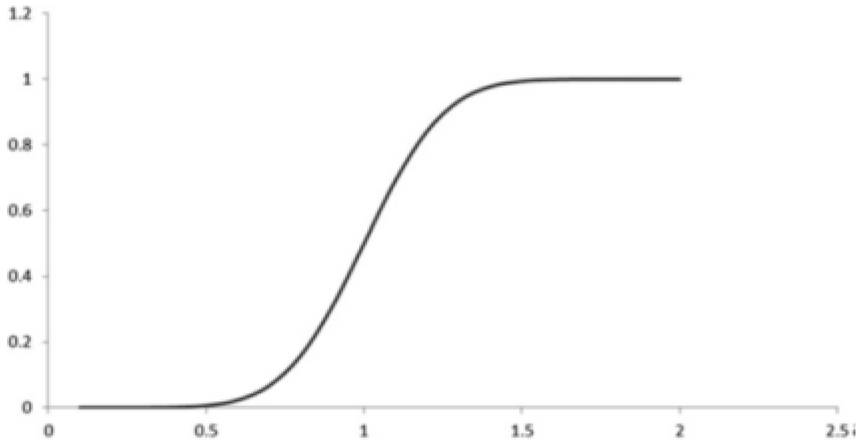


Figure 10: The Logistic function (Lorenz, 2015)

The Binary logistic regression is used in many fields, such as marketing, medicine and social sciences.

2.2.2 Performance Measure for Models

The process of discovering knowledge from data does not stop when the model is built, showing the need to know the performance of the model; to ensure its accuracy and efficiency before relying on it, Therefore, after building the model, it was necessary to find methods and algorithms to measure the performance of the model. Statistical methods are important in evaluating and examining models. There are many methods to test the performance of predictive models

This research will focus on performance measurement methods that are limited to models that choosing between two possibilities: Yes or No. The world is full of problems and models that involve decisions yes or no.

A simple example of predictive models whose decisions are limited to two possibilities Yes or No is the prediction model of pipe breaks, Figure (11) is represents the representation of the pipe breaks model. Any pipe will fall into one of the two classes and usually there is a break point between the two classes. This representation is a theoretical representation, but in practice, obtaining a definitive answer to the existence of the situation in one of the two classes is not possible.

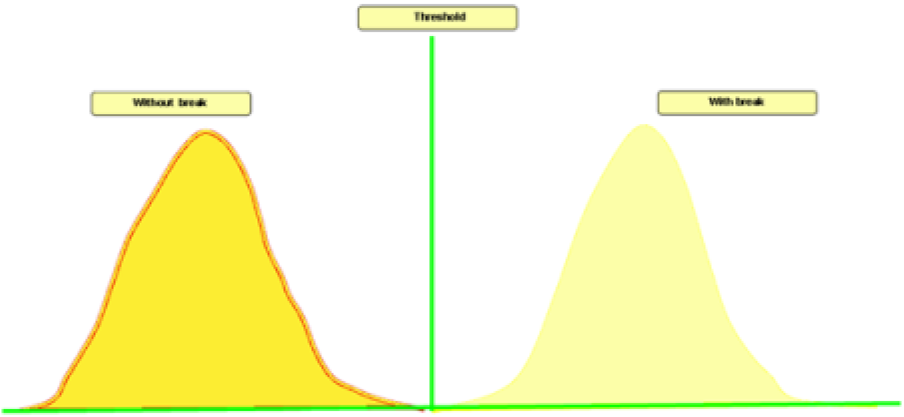


Figure 11: A theoretical representation of pipe breaks model.

In predictive models, there is usually an overlap between the two classes, where the model can predict the presence of a pipe in the first class but actually it exists in the other class as show in Figure (12).

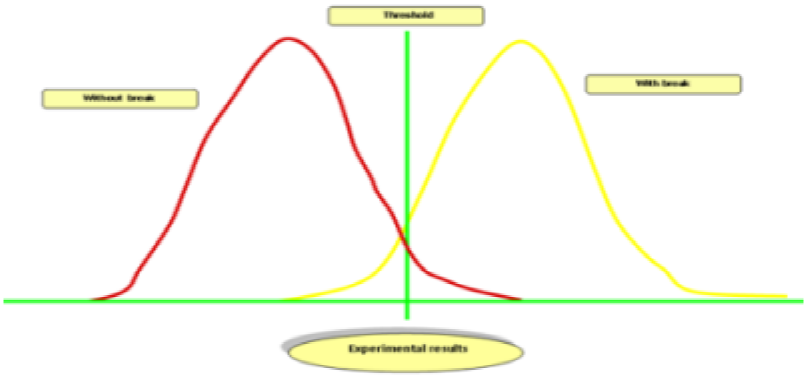


Figure 12: A practically representation of pipe breaks predictive model.

To represent predictive models in practice as show in Figure 13 , The overlap between the two distributions leads to different four cases, which are true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) (Fawcett, 2005).

True Positives (TP): These are cases in which we predicted Yes, and they do have the Yes, True Negatives (TN): These are cases in which we predicted No, and they really No, False Positives (FP): These are cases in which we predicted Yes, but they actually No and False Negatives (FN): These are cases in which we predicted No, but they actually Yes.

2.2.2.1 Confusion Matrix

The confusion matrix is usually used to represent the results of binary predictive models, also known as error matrix, which is a matrix contains information about actual and predicted classifications done by a classification system, as show in Table 1 It consists of two columns and two rows, where the rows represent the actual classes and are yes or no, and the columns represent the prediction classes of data that are yes or no. The intersection of rows and columns gives four cases are True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) (Sokolova & Lapalme, 2009).

The performance of the model can be measured and models performance can be compared with several methods, including the use of graphs such as ROC, Lift, Sensitivity, Precision, Hand and Cost curves. It is relied upon to calculate model performance metrics such as error rate, sensitivity (true positive rate), false discovery rate, miss rate, precision, specificity (true negative rate) and accuracy.

Table 1 The Confusion Matrix

		prediction	
		No (Negative)	Yes (Positive)
Actual	No (Negative)	True Negative (TN)	False Positive (FP)
	Yes (Positive)	False Negative (FN)	True Positive (TP)

The Error Rate is a rate of observations wrong in overall observations, which also known as Misclassification Rate. It can be calculated using the following equation

$$Error\ Rate = \frac{(FP+FN)}{\sum(TN+FP+FN+TP)}$$

The Sensitivity which also known as Recall or True Positive Rate is how often does it predict yes, when it's actually yes. It can be calculated using the following equation

$$Sensitivity = \frac{TP}{(TP+FP)}$$

The False Discovery Rate which display how often does it predict yes, when it's actually no. It can be calculated using the following equation

$$False\ Discovery\ Rate = \frac{FP}{(TN+FP)}$$

The Miss rate which display how often does it predict no, when it's actually yes. It can be calculated using the following equation

$$Miss\ Rate = \frac{FN}{(TP+FN)}$$

The Specificity which also known as true negative rate, is display how often does it predict no, when it's actually no. It can be calculated using the following equation

$$Specificity = \frac{TN}{(FP+TN)}$$

The Precision is when it predicts yes, how often is it correct. It can be calculated using the following equation

$$Precision = \frac{TP}{(FP+TP)}$$

The Accuracy is how often the classifier is correct in overall. It can be calculated using the following equation

$$Accuracy = \frac{(TP+TN)}{\sum(TN+FP+FN+TP)}$$

2.2.2.2 Cross Validation

When constructing a prediction model, it is necessary to estimate the performance of this model for a sample of data for future data. The dataset under study is usually divided into three sections: the training data set used to construct the model, the sample data set used to tune the model, and the testing set used to test the model's performance (Han & kamber, 2012).

the cross validation is one of the most important statistical techniques used to measure the performance accuracy of the model that was built and avoid the overlapping of testing datasets, and there are several types of the cross validation, the K- fold cross validation is the most famous and most important from it.

The K-fold cross validation algorithm divides the data into K equal data sets, then the model that was built is trained by the K-1 data sets, then the model is tested on the K data set, these steps are repeated k times, in each time the error is calculated, then the all errors are taken (Syed, 2011).

2.2.3 The Programming Language for Data Mining

Due to the fact that data mining techniques achieved success in several fields and spread widely, it was necessary to find applications and programming languages that support the application of these technologies to take different advantages of it.

But what are the best programming languages that can be used to build data mining models?

To answer of this question, two articles of web site are used (Piatetsky, 2014) (kdnuggets, 2012), which are displayed the languages are used for data mining techniques.

the languages can be used in data mining are R, Python, SQL, Java, SAS, C, C++, Matlab, Perl and Unix Shell/awk/sed languages and others.

In (Piatetsky, 2014) the top four programming languages that used for data mining are displayed, which are R, Python, SQL and SAS. It was found that about 91% of those who work in data mining use at least one of these languages in data mining tasks. It also turns out that the language that got the top one is R language.

2.2.3.1 R Language

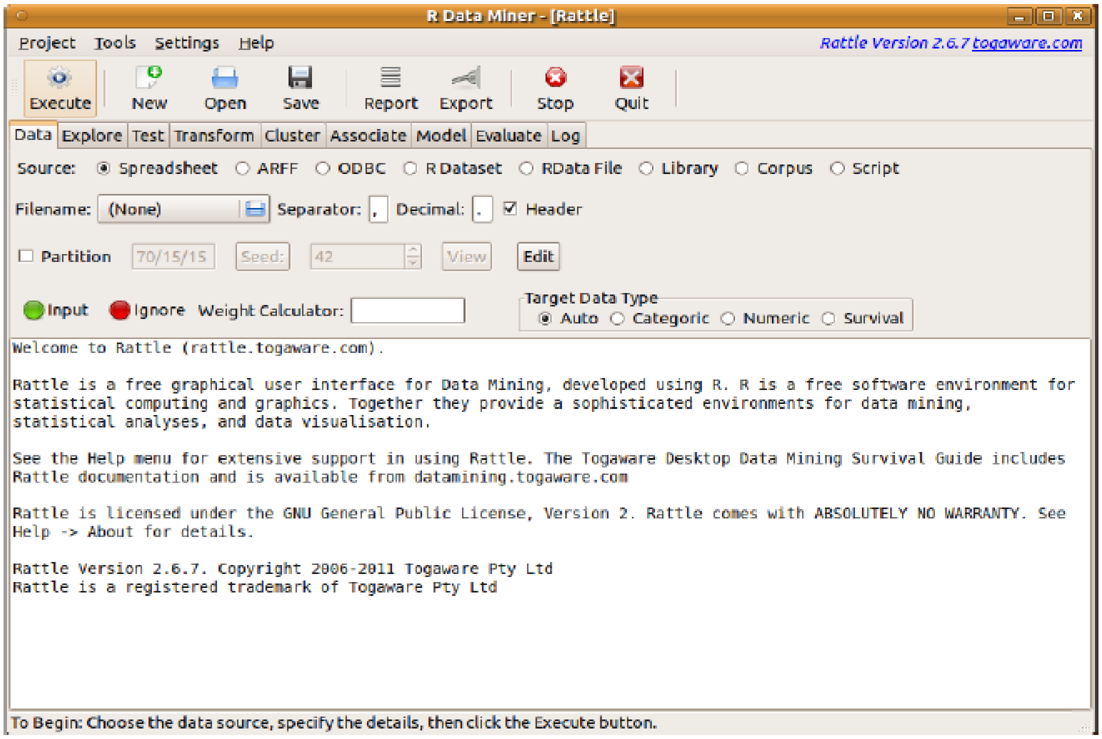
R language is one of the most widely used programming languages in the statistical analysis and data mining fields, which is an open source programming language for statistical computing and graphics (Zhao, R and Data Mining: Examples and Case Studies, 2011), (Venables & Smith, 2017) Publication of it is subject to GPL license, which led to an increase in the available resources for it on the Internet such as: e-books, educational courses, some ready-made programs written using them, which distinguishes the R language, which have available site with the possibility of downloading from the site directly, continuous updating of site and tools and performs all knowledge discovery in data (KDD) functions rather than a specific function. R language supports most common operating systems such as Apple, Windows and Linux. Its setup is easy and uncomplicated. It also allows building the predictive and descriptive models using different methods and algorithms. The use of the R language of the data analysis is not limited to text, but rather allows the representation of the data graphically.

Since R is open source this has made it a constantly evolving language, as the new methods and algorithms are being developed and introduced to it, which are formed additions to this language.

These additions are usually done by researchers and graduate students, who are motivated by the ease of building additions to this language, which provides many tools free and rapidly, much better than the high-priced commercial alternatives that are controlled by the manufacturers.

These additions are available as a hundreds of ready packages, which supports many data mining tasks. Some important packages in the R language are: rpart, kernlab, e1071, random forest, rattle, party, arules and grphics and many other packages (Team, 2016).

The Rattle package is important in this research. Rattle term in R language is a shortcut to the R Analytical Tool to Learn Easily that is the data mining interface for R language. Rattle is available as a ready-to-use package for the R language و it provides a simple and easy graphical interface for users of R language as shown in Figure (13), especially beginner users.


R Data Miner - [Rattle]
Project Tools Settings Help
Rattle Version 2.6.7 togaware.com
Execute New Open Save Report Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Input Ignore

Source: ◉ Spreadsheet ○ ARFF ○ ODBC ○ R Dataset ○ RData File ○ Library ○ Corpus ○ Script

Filename: (None) Separator: , Decimal: . ☑ Header

☐ Partition 70/15/15 Seed: 42 View Edit

Weight Calculator:

Target Data Type ◉ Auto ○ Categoric ○ Numeric ○ Survival

```
Welcome to Rattle (rattle.togaware.com).

Rattle is a free graphical user interface for Data Mining, developed using R. R is a free software environment for
statistical computing and graphics. Together they provide a sophisticated environments for data mining,
statistical analyses, and data visualisation.

See the Help menu for extensive support in using Rattle. The Togaware Desktop Data Mining Survival Guide includes
Rattle documentation and is available from datamining.togaware.com

Rattle is licensed under the GNU General Public License, Version 2. Rattle comes with ABSOLUTELY NO WARRANTY. See
Help -> About for details.

Rattle Version 2.6.7. Copyright 2006-2011 Togaware Pty Ltd
Rattle is a registered trademark of Togaware Pty Ltd
```

To Begin: Choose the data source, specify the details, then click the Execute button.

Figure 13: rattle interface.

Rattle package is used to provide the ability to analyze data and build strong descriptive and predictive models by using many powerful tools and algorithms available in the R language, in addition to the possibility of evaluating the models that have been built, until the pattern is completed and its deployment without the need to know and use the written commands in the R language (Williams, 2009).

R language and this packages with a thousands of packages is available on CRAN web site as a free software environment in addition to other many websites , which can be extended very easily by many of developers (Team, 2016)

2.3 Related Works

This section introduces nine different research papers that present mathematical and statistical solutions to the perennial problems in management of water distribution systems. Problems faced by utility managers include; lack of deterministic models to predict pipe bursts, leaks and costs involved. From data collected, it is obvious that current manual techniques for monitoring, managing and understanding water management systems are inefficient, outdated and costly. In the below papers, researchers present different techniques that can be applied to improve the entire water management value chain, from data collection to analysis and interpretation. Solutions presented include regression models, data mining algorithms, stochastic processes and genetic programming techniques.

Paper (Giustolisi, Savic, & Laucelli, 2004) introduced EPR (Evolutionary Polynomial Regression), a better technique that can be used to predict occurrence of pipe bursts. EPR makes use of symbolic data analysis techniques and has been tested using UK water distribution system data sets. Apart from predicting occurrence of pipe bursts, EPR also investigates occurrence and frequency of factors that contribute to pipe bursts such as age, diameter and length of pipes. Generally, results from use of EPR models showed a big positive correlation between age and length of pipe to pipe burst. There was an inverse relationship between pipe diameter and pipe bursts. EPR presents a reliable and systematic way to predict and monitor occurrence of pipe bursts.

Researchers in (Mashford, Silva, Marney, & Burn, 2009) seek to investigate pipe leakages in South Eastern Melbourne, which showed practically applicable methods for locating and determining location of pipe leaks in water systems. This process simply requires monitoring of pressures at specific location in a water system and use of SVMs. Data required to train SVMs is in turn obtained from EPANET systems. Results show that accuracy of prediction of leak location majorly depends on spatial resolution of used data. However, for prediction of leaks in water systems to be possible pressure sensors need to be installed in different sections of a water system.

Paper (Bubtiena, ElShafie, & Jaafar, 2011) introduced much needed and adapTable model improvement techniques that can be widely applied in current regression models for pipe bursts predictions, through improve accuracy of existing non-linear multiple regression models, this research using real life water distribution system. The researchers showed practical ways through which system engineers can improve predictability pipe breakages.

Study (Xu, Chen, Li , & Ma, 2011) assessed different statistical models that can be used to determine pipe brakes in water systems. Two deterministic statistical data techniques are used for this analysis; GP (Genetic Program Mining) and EPR (Evolutionary Polynomial Regression). This study was based in Beijing city which was used as case study. Researchers in this paper argue that there is need to sort and order the historical pipe break data available complex data available. The data can be ordered according to the years, quantity and magnitude. This paper successfully presented a data sorting procedures that can be used for future pipe breaks data collection and storage.

Research in (Martins, Leitão, & Amado, Comparative Study of Three Stochastic Models for Prediction of Pipe Failures in Water Supply Systems, 2011) makes use of different stochastic processes like the poisson, Weibull and yule process to predict failure in Portuguese water systems. After applying the three models it was however discovered that Weibull process showed the best process results by correctly predicting occurrence of water system failures. Results from this research showed that pipes that have a history of failure have a higher likelihood of failing again, this because most pipes become more vulnerable after repairs.

In study (Wang, Dong, Wang, Tang, & Yao, 2013) the researchers use data mining and statistical techniques to design and schedule maintenance prevention techniques for water systems. Aim of the research was to reduce the costs and improve the accuracy when it comes to water system maintenance and improvement. The research draws data from New York City. The research has shown that by use of algorithms like RankBoost.B empirical studies can be analysed to allow schedule maintenance of water pipes which in turn prevents pipe breakages.

Study (Xu, Chen, Ma, & Blanckaert, 2013) used the Beijing water system as a case study and propose statistical solutions to the problem of water pipes breakages. The prediction model was then developed using genetic programming. This research successfully determines and connects pipe breaks and replacement costs using real life data collected in Beijing city. The model uses genetic programming to predict break rate at a considerable performance rate of 0.814. The model also calculates the approximate replacement time for pipes given their diameters and length. Economic equations introduced in the study give an estimate of the cost of future replacements and repairs which is not only crucial but is also cost effective to utility managers. The study recommends assimilation of new data in the model to improve its efficiency.

The (Francis, Guikema, & Henneman, 2014) discusses application of Bayesian inference to forecast the pipe breaks in mid-Atlantic US water system. The result, application of BBN in prediction of pipe breaks presents and important discovery for the general use of statistical inference techniques and models in understanding complex systems. Though the results show that this is a complex procedure, BBN study not only presents a significant knowledge base, it can also present a framework for the adoption of data collection standards in the water management industry.

In (Ghorbanian, Guo, & Karney, 2016)the researchers make use of historical data to design a probabilistic model for determining occurrence of pipe breaks and bursts. This study was based in Ontario, Canada. Monte Carlo simulation is used to estimate rate of expected pipe breaks and reduce costs incurred in repairs and rehabilitation. The results, when the pressures reduced will decrease the pipes broken in system.

Chapter Three
Model Development

3.1 Introduction

The purpose of this research is to use the data available in the records of the Studies and Maintenance departments at the Water and Sanitation Department in the Nablus Municipality in to solving the problem of water loss from broken pipes through the use of data mining techniques in order to design a predictive model for broken pipes. To achieve this, we must understand the general perception of the research which is displayed in Figure (14).

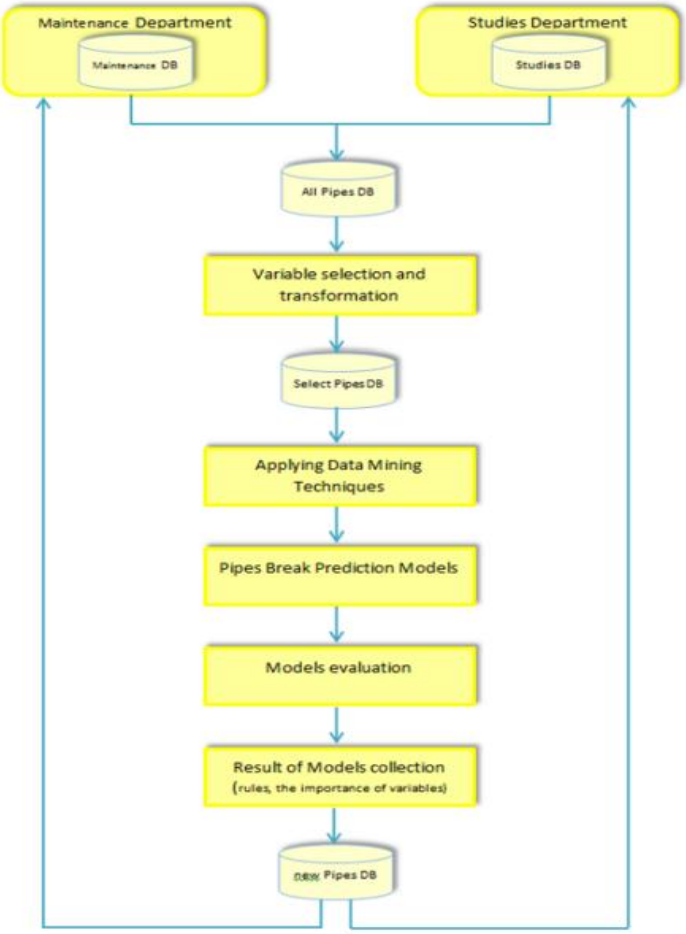

Maintenance Department
Maintenance DB
Studies Department
Studies DB
All Pipes DB
Variable selection and transformation
Select Pipes DB
Applying Data Mining Techniques
Pipes Break Prediction Models
Models evaluation
Result of Models collection
(rules, the importance of variables)
new Pipes DB

Figure 14: General perception of the research.

This research adopts the inductive research method in which the researcher begins with the collection of data and observations and then moves forward to generalizations; the generalizations in this case are related to pipe break prediction model.

To achieve the objective of this research, the process of developing the model must be well thought and sequenced. Figure 15 displays a sequence diagram, which presents the steps of data collection and analysis used to build models, implement them and eventually assess them in order to choose the best. In the following section, the used research methodology is presented in details.

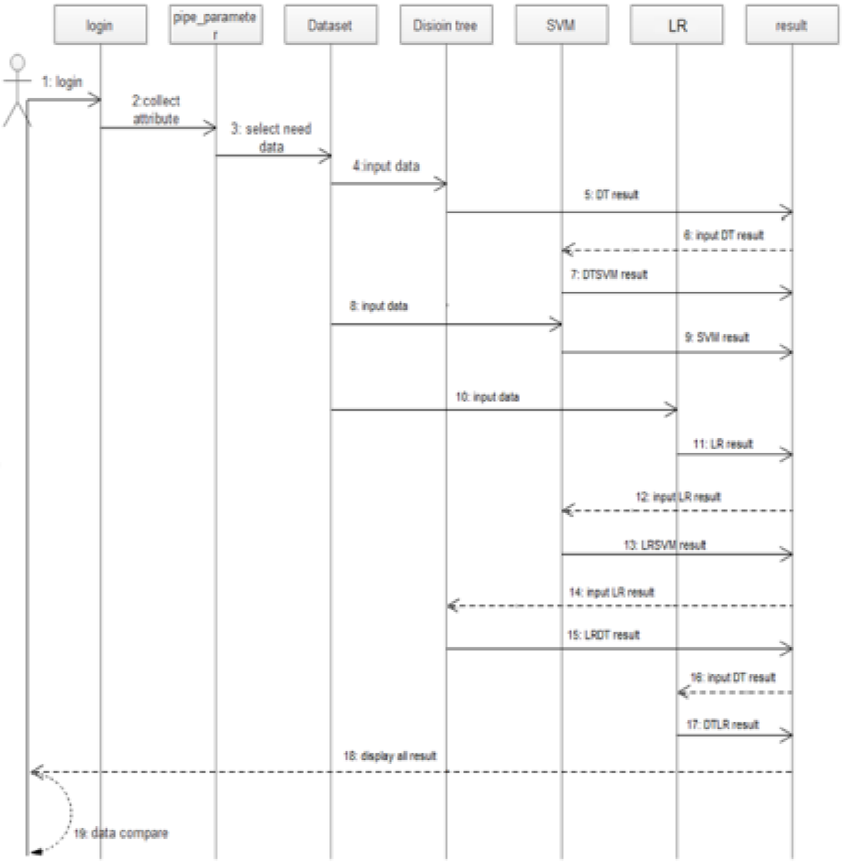


Figure 15: The sequence diagram of the search

3.2 Research Methodology

The research methodology begins from understanding the know data related to the existing system. The next step is data collection and processing in terms of size, shape and content; the data in this study is to be collected from Nablus Municipality departments. The third step would be the selection of the variables related to the prediction model. The final step is data transformation to a suitable form consistent with model construction.

Since building and evaluating models is an important step to achieve the objective of research; so that seven binary classification models are building, evaluating and comparing performance of it. The steps of research methodology are displayed in the flowchart in Figure (16).

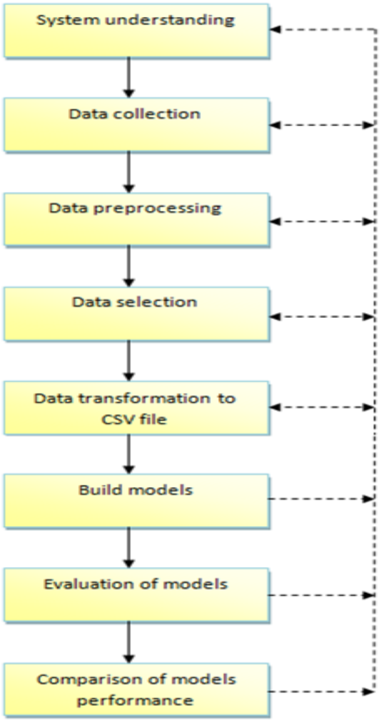


Figure 16: Flowchart of research methodology.

The first step which is system understanding was explained in detail in the first chapter, in the following section the other steps are explained in detail.

3.2.1. Data Collection

The information related to 4810 main pipes in the network was obtained from the Studies Department. The data mainly pertained to pipe ID (A distinguishing serial number of the pipe that), pipe material, installation date, pipe diameter, pipe length, pipe type (its distribution "1 or 2"), pipe subtype, pipe pressure zone (the code of water pressure in pipe), pipe pressure area (the name of area that has the water pressure in the pipe), and the layer of the pipe. As shown in Table 7.

Table 2: Variable of main pipes from studies department.

Features	Values
Pipe Material	Poly Propline, HDPE, Poly Vinyl Chloride, Steel, Ductile Iron, Galvanized steel
Year laid	From 1957 to 2017
Diameters	From 2 to 12inches
Lengths	Total 298.20 km
Shape Length	From 0.010202 to 555.198138
pressure zone	Asker_Camps, C1, E0.1, E0.2, E0.3, Ein_camp, Gasial, Hurush_line, NE1, NE2, NE4, NW0, NW1+NW2a, NW3+NW2b, Out Nablus, Rafedia, S2, S3,S4, S5, SE1, SE2, SE3, W0, W1, W-1, W2a, W2b, W3, W4, Worash_Gravity, Zawata.
Type	Dis_1, Dis_2.
Subtype	From 1 to 20

The data obtained from the Studies Department is composed of two Excel files, the first one shows type distribution 1, and the second one displays type distribution 2.

Information recorded for pipes maintenance in the network was obtained from maintenance department, which include data of all pipes had maintained from year 2010 to 2016.

Those records containing data of pipes maintenance related to maintenance type, type of pipe that main pipe or not, in addition to the data for the pipe such as pipe materials, pipe diameters etc.

3.2.2. Data Preprocessing

After completing the data collection, at first the two files from Studies Department are merged in one Excel file and then some operations are performed operation on file to be clean and normal.

When looking at the data file, it was found that part of it was measured by the inch unit while the other part was measured by the Millimeters unit, a lack of data was shown about knowing the pipe installed year for some pipes.

After completing the data collection, at first the two files from Studies Department are merged in one Excel file and then some operations are performed operation on file to be clean and normal.

When looking at the data file, it was found that part of it was measured by the inch unit while the other part was measured by the Millimeters unit and a lack of data was shown about knowing the pipe installed year for some pipes.

3.2.2.1. Normalization the Diameter of the Pipes

Part of the diameter of the pipes was measured by the inch unit while the other part was measured by the Millimeters unit, so must work normalization for it, to normalization the diameter of the pipes an algorithm was used to make all the diameters of the pipes in millimeter. Find Missing Data for Pipe Install Year

A missing of data was shown about knowing the pipe installed year for some pipes. By asking the engineers working in the studies department in the water and sanitation department in the Nablus Municipality about the lost installed year of pipes data, it was found that the missing data belong to the pipes that were installed before 1990. Therefore, the data for the pipes that were installed before 1990 are extracted. Then an apriori algorithm was applied on the data from the pipes that were installed before 1990 to get the lost years and get rid of the missing data. as shown below pseudo code for find missing year.

After getting rid of the missing data, the pipe age of pipe was calculate by subtraction the install year from the research year, and insert it to file as new column. The maintenance data for broken main pipes from years 2010 to 2016 were selected that are interest in this research

In view of the data some difference was observed in the names of some variables, so the names of the variables are unified before the two data are merged together. Then this data are merged.

3.2.3. Data Selection

After the data are cleaned and integrated it becomes ready to build models but must be selected the variables relevant to the research task and problem. so that the selection data process is performed. Then variables are reformulated to become well and suiTable to build models. the Variables selected and reformulated to build the prediction models are The pipe ID (ObjectID), pressure zone for pipe (pressure_zone) ,subtype of pipe (subtype), pipe materials (material),layer of pipe (layer),type of pipe (type) Shape Length of pipe (shape_length),pipe diameters (Diam), pipe break , and age of pipe (pipe_age) as shown in Table (3)

Table 3: The variable selected.

Variables selected	Values
ObjectID	From 1 to 5096
Pipe_age	From 1 to 63 years
Pipe Material	Poly Propline, HDPE, Poly Vinyl Chloride, Steel, Ductile Iron, Galvanized steel
Diameters	From 2 to 12inches
pressure zone	Asker_Camps, C1, E0.1, E0.2, E0.3, Ein_camp, Gasial, Hurush_line, NE1, NE2, NE4, NW0, NW1+NW2a, NW3+NW2b, Out Nablus, Rafedia, S2, S3,S4, S5, SE1, SE2, SE3, W0, W1, W-1, W2a, W2b, W3, W4, Worash_Gravity, Zawata.
Type	Dis_1, Dis_2.
Pipe break	Yes or No

3.2.3.1. Data Loading

R language is used to loading dataset, which is previously collected and processed, preface to build the model.

data is loaded directly from a comma separated value (CSV) file, which is contains data on the pipeline network in the municipality of Nablus, contains 4810 records of pipelines and maintenance carried out on them. After the data are loaded the next step is to define the variables in the file and determine the type and role of each of them in building the model. The variables Pressure Zone, Subtype, Material, Type, Diam and pipe.age are selected as input variables, which are will be used to predict to target variable

Input variables are divided into two types numeric and categoric variables. the numeric variables are Subtype, Shape_Length, Diam and pipe.age. The categoric variables are Pressure_Zone, material and type.

The variable pipe.break is selected as target of models and the Objectid as ident for records, which is ignored when modeling but it is used when scoring dataset to ident observations in scoring file.

3.2.3.2. Data Exploring

The better understanding of the available data set is important to know what the data look like before building any prediction models for water pipes break. So the exploring data will be necessary to do.

To exploring data, the each one of the input variables are studied to analysis the distribution of it by pipe break, and understand the frequency and density of each variable by pipe break. the Graphical tool are used to achieve that. The Graphical tool is a good to help us to understand the data's characteristics and distribution of it by pipe break.

This section display the histogram plot for numeric features, which is a common way of presenting data because it provide useful and quick graphical view of the distribution of the data, and Bar blot for categorical features, which like histogram, that uses to show the frequency of each of the possible values of the categorical variables.

3.2.3.2.1. Distribution of Diameter by Pipe Break

The diameter of the pipes in the research dataset is ranged from 50 to 300 millimeter (mm). but does the pipe diameter have any effect on pipes break? That is, is the occurrence of broken pipes concentrated at a particular pipe diameter? To answer of these questions must show the distribution of pipe diameters by pipe break. So that the Figure (17) display the distribution of pipe diameters by pipe break, where it shows that the broken pipe is concentrated in pipes with a diameter ranging from 60 to 90. While the frequency of pipes break is very small for pipes with a diameter between 200 to 300 millimeters.

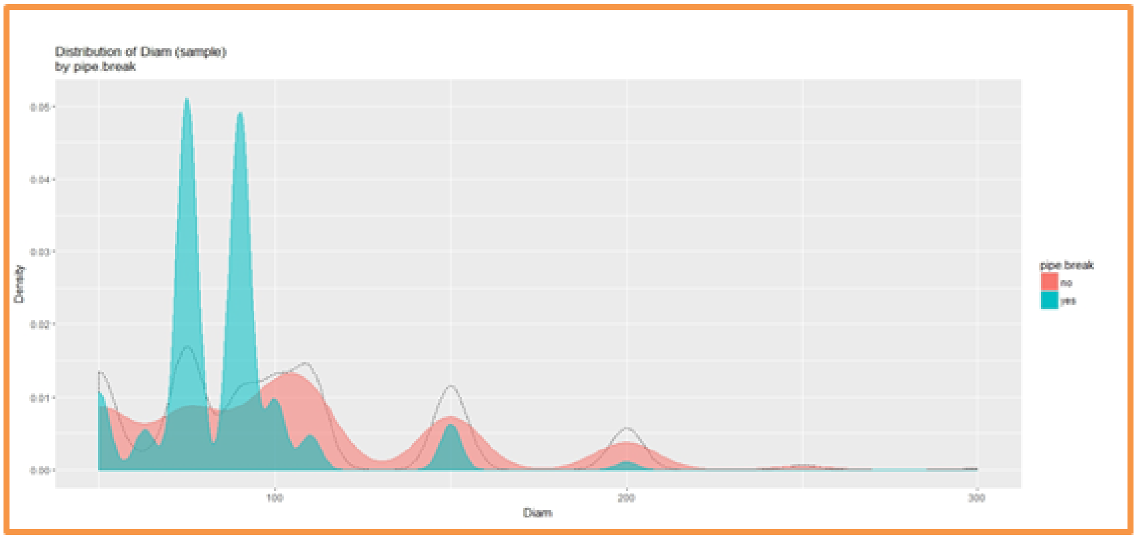


Figure 17: Histogram plot for distribution of diameter by pipe break.

This means that there is an inverse relationship between pipe break and pipe diameter, as the frequency of pipe breaking increases as the diameter of the pipe decreases.

3.2.3.2.2. Distribution of Subtype by Pipe Break

The subtype of pipe in the research dataset takes values from 1 to 20. The frequency of pipe break in each one of subtype values are studied to know the distribution of subtype by pipe break and the relation between subtype by pipe break. as shown in Figure (18) The biggest frequency of the pipe break occurs when the subtype value equal ten. Also when the subtype is equal to eight, the density of the pipe break is close to the previous one.

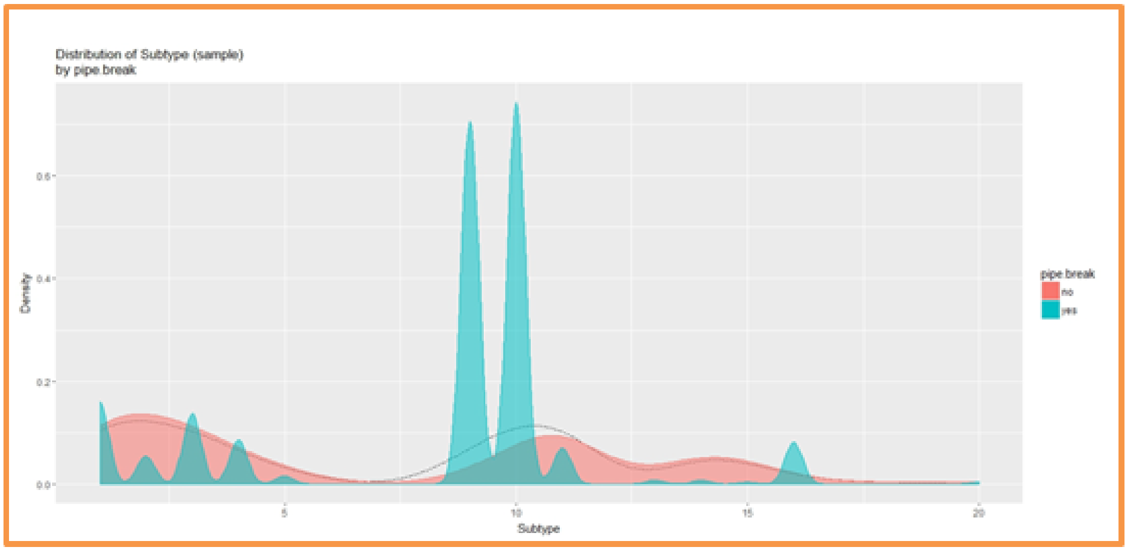


Figure 18: Histogram plot for distribution of subtype by pipe break.

However, it is observed that the frequency of the break of the pipes is almost distributed to all subtype values, which means that it is difficult to determine the relationship between the subtype and the pipes break accurately.

3.2.3.2.3. **Distribution of Pipe Age by Pipe Break**

In reference to the data set that was prepared for the building of the prediction model, especially the variable age of the pipe, it is noted that the age of the main pipes in the water distribution network in the Nablus municipality ranges from one year to three sixty years.

A study was conducted on distribution of pipe age by pipe break. The result was as shown in Figure (19), Which shows that the density of broken pipes are concentrated in pipes that are its age more than thirty years old compared to the number of it. But it is remarkable that there is a high frequency of pipe break in pipes that age are 12 years old on average, Which are the pipes that were installed between 2007 and 2012.

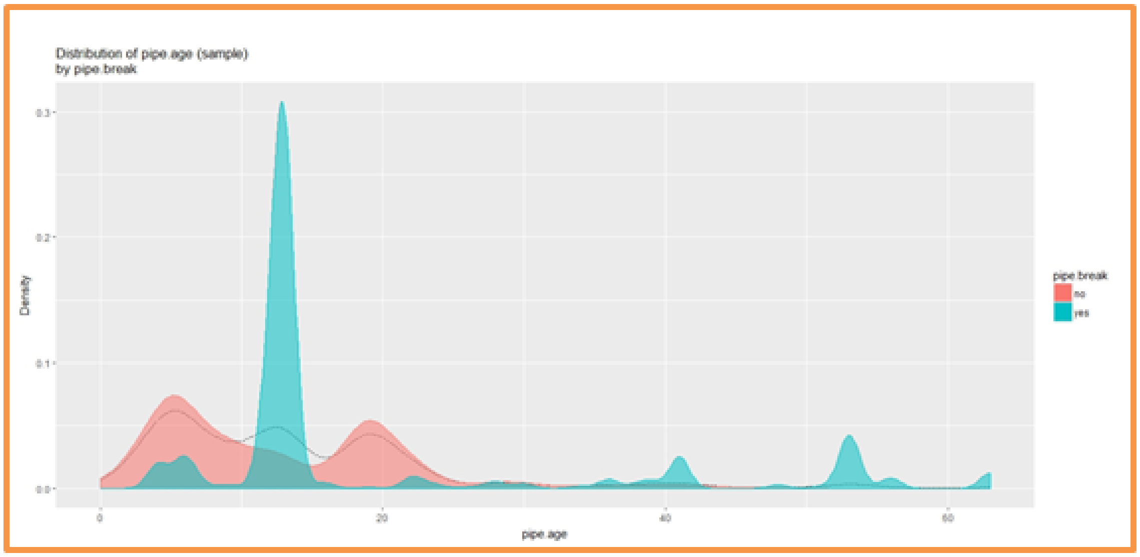


Figure 19: Histogram plot for distribution of pipe age by pipe break.

3.2.3.2.4. Distribution of Shape Length by Pipe Break

The histogram plot for distribution of shape length by pipe break is drawn to study the dinsety and frequency of the pipes break in shape length values as shown in Figure (20) ,Which shows that the pipes with break are almost intertwined with the pipes without break

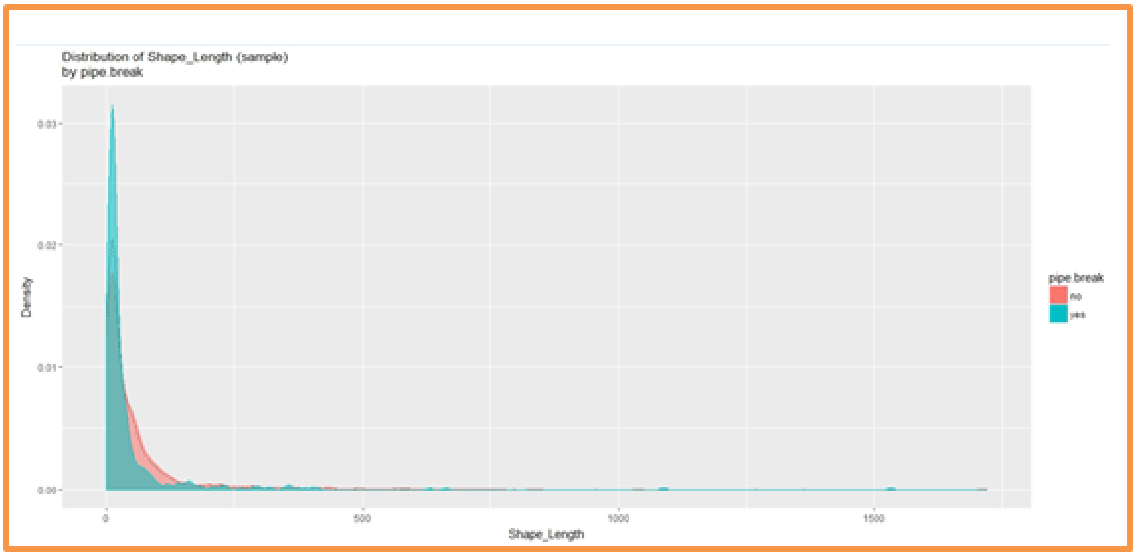


Figure 20: Histogram plot for distribution of shape length by pipe break.

3.2.3.2.5. Distribution of Material by Pipe Break

The material variable is one of categorical variables in research dataset, which have six different values, that is PP (It is an abbreviation for Poly Propline pipe), HDPE (It is an abbreviation for High-density polyethylene pipe), ST (It is an abbreviation for Steel pipe), GS (It is an abbreviation for Galvanized steel pipe), PVC (It is an abbreviation for Poly Vinyl Chloride pipe) and DI (It is an abbreviation for Ductile Iron pipe).

To understand the frequency of the pipes break in material variable values the distribution of material by pipe break is drawn by Bar blot as shown in Figure (21), because the histogram plot can't be used to draw The categorical variables, Which shows that there is a significant difference in the frequency of the pipes used in the water distribution network in the municipality of Nablus, which the order of the pipes by frequency from the most frequency to the least as follows: Galvanized steel pipe, High-density polyethylene pipe, Poly Propline pipe, Ductile Iron pipe, Steel pipe and Poly Vinyl Chloride. While they are arranged by the frequency or density of the broken pipes in it, relative to their frequency from the most to the lowest fraction as follows: Poly Propline pipe, Steel pipe, High-density polyethylene pipe, Galvanized steel pipe, Ductile Iron pipe and Poly Vinyl Chloride

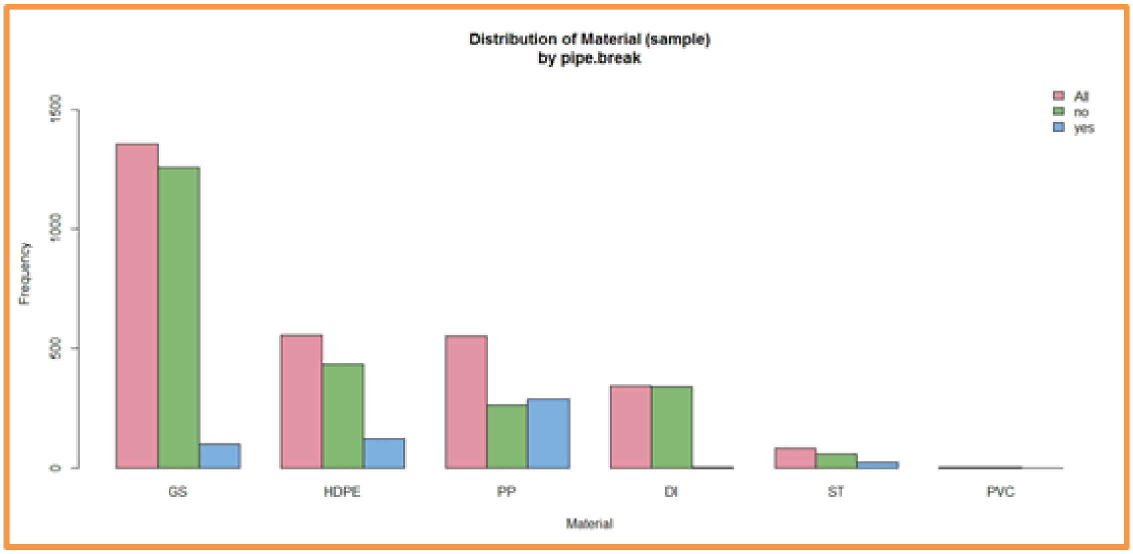


Figure 21: the Bar plot for distribution of material by pipe break.

3.2.3.2.6. Distribution of Pressure Zone by Pipe Break

The pipes in the water distribution network in the municipality of Nablus are distributed into 32 different water pressure zones,

To determine whether the pressure value was related to the fracture of the pipes, the distribution of the pressure zones by pipe break is studied by the Bar blot graph as shown in Figure (22).

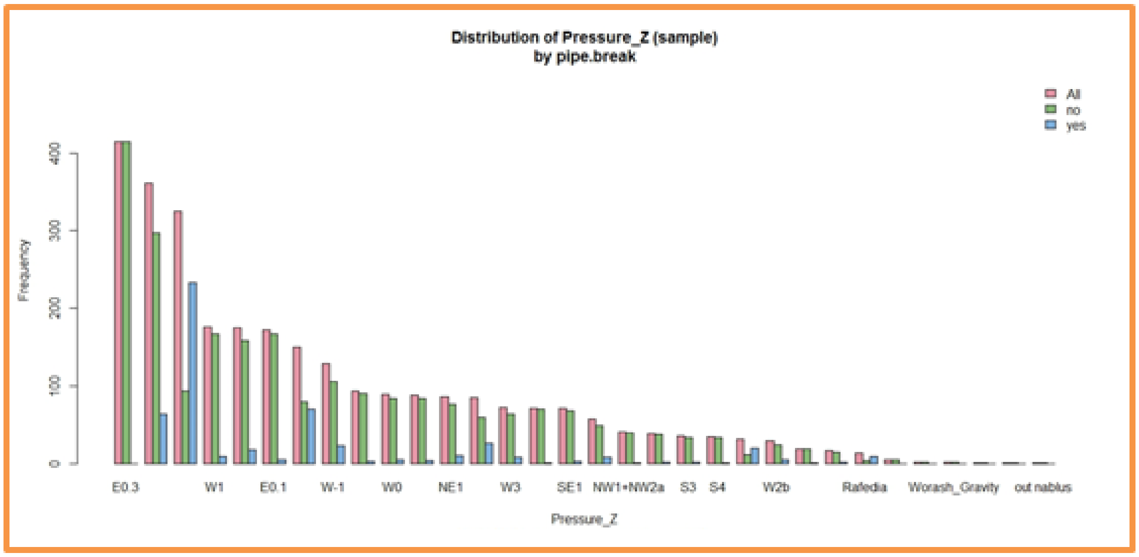


Figure 22: the Bar Plot for distribution of pressure zone by pipe break.

3.2.3.2.7. Distribution of Type by Pipe Break

The pipes in the water distribution network in Nablus Municipality have one of two types are dis_1, which represent the first part of the water distribution network and dis_2, which represent the second part of the water distribution network according to data from studies Department. To understand the frequency of the pipes break in the type variable the distribution of type by pipe break is explained by the Bar blot as show in Figure (23), which is shown that the density of the pipes break in the dis_1 type is much greater than in dis_2 type.

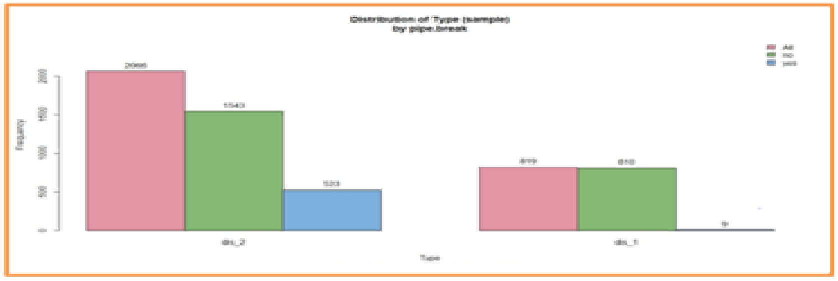


Figure 23: Bar blot for distribution of type by pipe break.

3.2.3.3. Models Training, Validation and Testing

Three datasets are need when the prediction model will build; where the first dataset will use to build the model, the second will use to tune the model building to improve its performance and the last dataset to evaluate the model performance.

To obtain data sets and build a prediction model our dataset was divided into three parts are: training dataset which will use to build the model, validation dataset (known as a design dataset) because it assist in the design of the model and tune it, and testing dataset (known as evaluation dataset) to evaluate the performance of model.

This research is adopted Prof. Andrew Ng's recommendation in his online machine learning course to divide the dataset into: 60% as Training data, 20% as validation data and 20% as testing data (Ng, 2016)

After loading data the first step to divide data, which a seed for generated sequence of random number to divide the dataset is selected, The seed is putted the same number for each building models to obtaining of the same sample .The number of seed her is 46. Then the observations of our dataset are taken and divide into three data sets. Now the data are ready to build the models

3.2.3.4. Implementation of Building Models

After the data sets became ready through dataset partition, it became possible to build models to predict the breaks of the pipes. As shown in Figure (24).

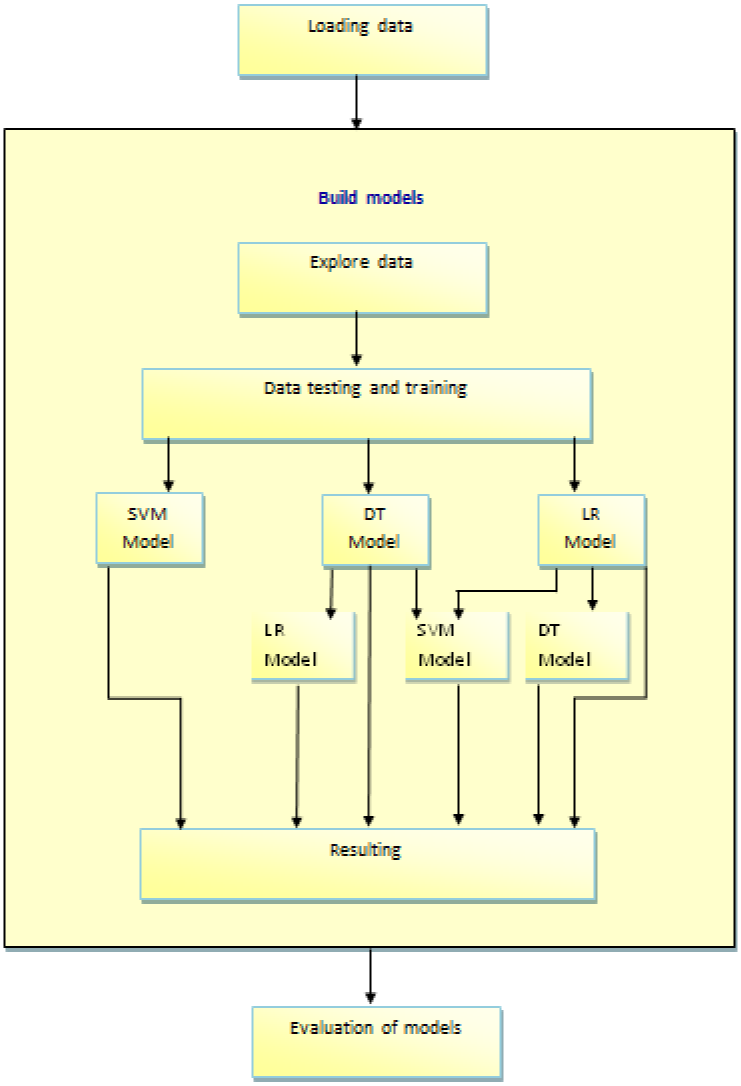


Figure 24: Building models.

In this section, seven classification models have been built to predict pipes break which are depending on using the DT, SVM and LR algorithms.

3.2.3.4.1. Decision Tree Model (DT)

The first model was built using the decision tree algorithm the Rattle in R language are used to build the decision tree classification model. The traditional decision tree algorithm are used to build the decision tree model, which is used the 'rpart' package, that is provides the reliably and very robust implementation of the classification algorithm. The rpart package is a ready-made software package available for R language, which used for the purpose of building and representing classification models using a binary decision tree.

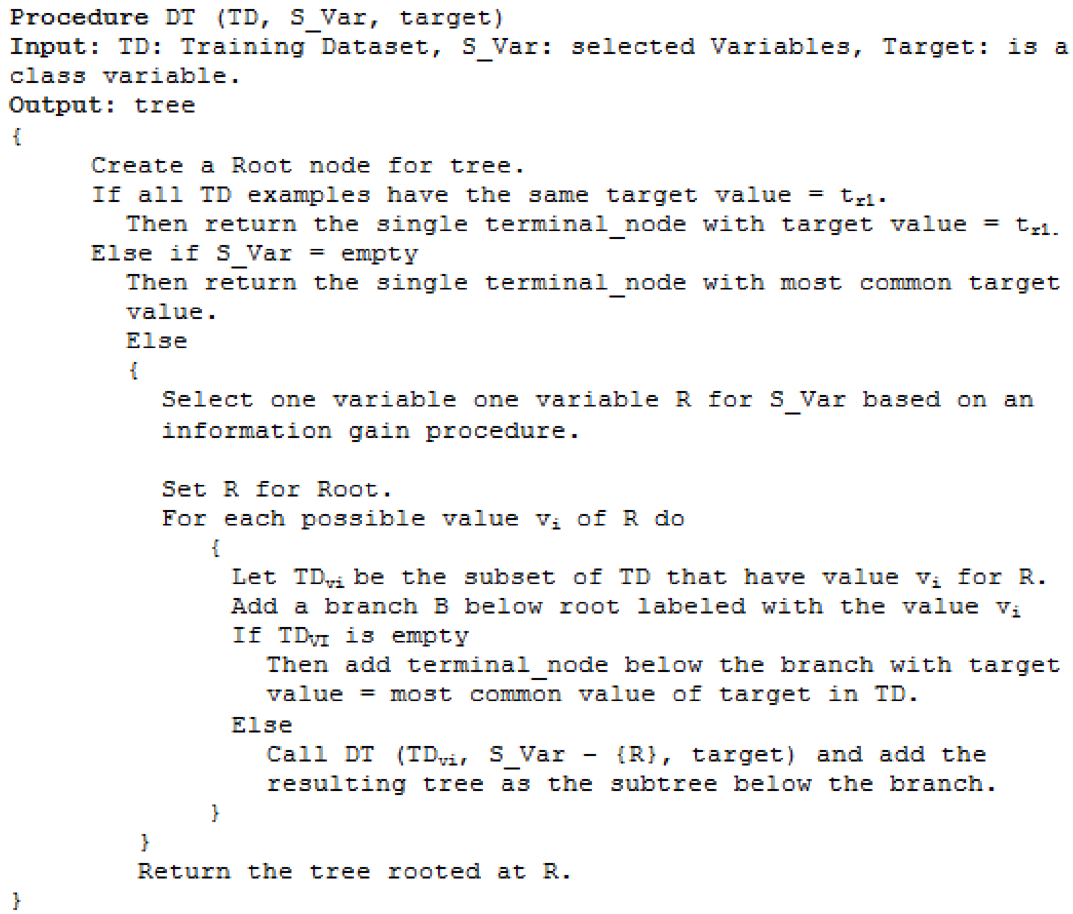
It includes tools to build the model, graphically represent it and give a summary of the final results, which are easy to use and understand (Strobl, Malley, & Tutz, 2009) (Therneau, Atkinson, & Foundation, 2017).

There are a number of parameters whose values must be determined to build a Decision tree using this package, which is Min split, Min Bucket, Max Depth, Complexity, Priors and Loss matrix.

The Min split parameter represents the minimum split of observations in the node. It shows the minimum number of observations, which must be in any node in the tree until they are broken into another nodes, that is here 20 observations. The Min Bucket parameter is the minimum number of observations in any leaf node, and its value usually one third of the Min split parameter that is here 7 observations.

The Complexity parameter is a useful parameter in the tree pruning where it is used to adjust the size of the tree; this research takes the default value of the complexity in the rattle. The Max Depth parameter represents the depth of the tree, and is the largest number of levels that allow the tree to extend to it, that is here 30 levels.

The Priors parameter uses to specify percentages and priorities for each class in the classification model. The Loss matrix parameter The Loss matrix parameter uses to give weight to different error types, which are false positive and false negative in binary classification. In the last two parameters values are not specified, which means giving the two classes the same effectiveness in building the model, as well as the resulting error types. The DT algorithm Pseudocode is display in Figure (25).



```
Procedure DT (TD, S_Var, target)
Input: TD: Training Dataset, S_Var: selected Variables, Target: is a
class variable.
Output: tree
{
      Create a Root node for tree.
      If all TD examples have the same target value = $t_{r1}$.
        Then return the single terminal_node with target value = $t_{r1}$.
      Else if S_Var = empty
        Then return the single terminal_node with most common target
        value.
        Else
        {
          Select one variable one variable R for S_Var based on an
          information gain procedure.

          Set R for Root.
          For each possible value $v_i$ of R do
              {
               Let $TD_{vi}$ be the subset of TD that have value $v_i$ for R.
               Add a branch B below root labeled with the value $v_i$
               If $TD_{VI}$ is empty
                 Then add terminal_node below the branch with target
                 value = most common value of target in TD.
               Else
                 Call DT ($TD_{vi}$, S_Var - {R}, target) and add the
                 resulting tree as the subtree below the branch.
              }
         }
         Return the tree rooted at R.
}
```

Figure 25: The DT algorithm Pseudocode.

3.2.3.4.2. Support Vector Machine Model (SVM)

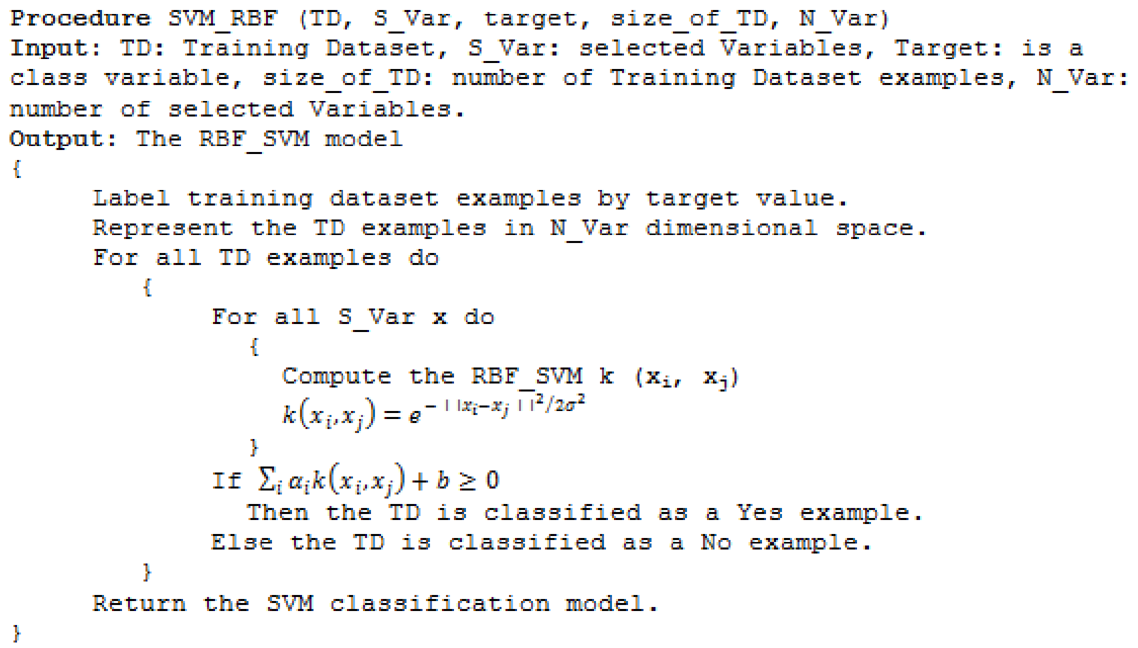
The Support Vector Machine (SVM) model is the second model is built in this research. Rattel supports the implementation of the SVM algorithm through the kernlab package that provides a ksvm function. The ksvm function can be used for different modeling tasks.

The kernlab package is one of the most important software packages available in the R language, which provides the ability to build classification models using SVM algorithms, which provides many basic kernel functions for the language R users, and It contains of many functions, which The ksvm() function is one of the most important of it.

The ksvm() function is the implementation of support vector machine (SVM) by kernlab, that provides a matrix interface with a standard formula interface (Karatzoglou, Meyer, & Horni, 2006).

The purpose of using the SVM algorithm in this search is to build a classification model, which means the use of C-SVC formulation in ksvm function, which is used to build svm classification model. the radial basis function kernel is used to represent our data into a multi-dimensional space, which using a Gaussian type of function (that is works on the principle of do dot product between two vectors) by using the automatic sigma value based on the data and parameter cost equal one to get a few the points near the decision boundary .

The last parameter is tuned that is the class probabilities parameter, which is used to determine if the model will calculate the class probability or not, that in this research true to calculate the class probability. The SVM-RBF algorithm Pseudocode is display in Figure (26)



```
Procedure SVM_RBF (TD, S_Var, target, size_of_TD, N_Var)
Input: TD: Training Dataset, S_Var: selected Variables, Target: is a
class variable, size_of_TD: number of Training Dataset examples, N_Var:
number of selected Variables.
Output: The RBF_SVM model
{
      Label training dataset examples by target value.
      Represent the TD examples in N_Var dimensional space.
      For all TD examples do
          {
               For all S_Var x do
                  {
                     Compute the RBF_SVM k ($x_i$, $x_j$)
                     $k(x_i,x_j) = e^{-||x_i-x_j||^2/2\sigma^2}$
                  }
               If $\sum_i \alpha_i k(x_i,x_j) + b \geq 0$
                  Then the TD is classified as a Yes example.
               Else the TD is classified as a No example.
          }
      Return the SVM classification model.
}
```

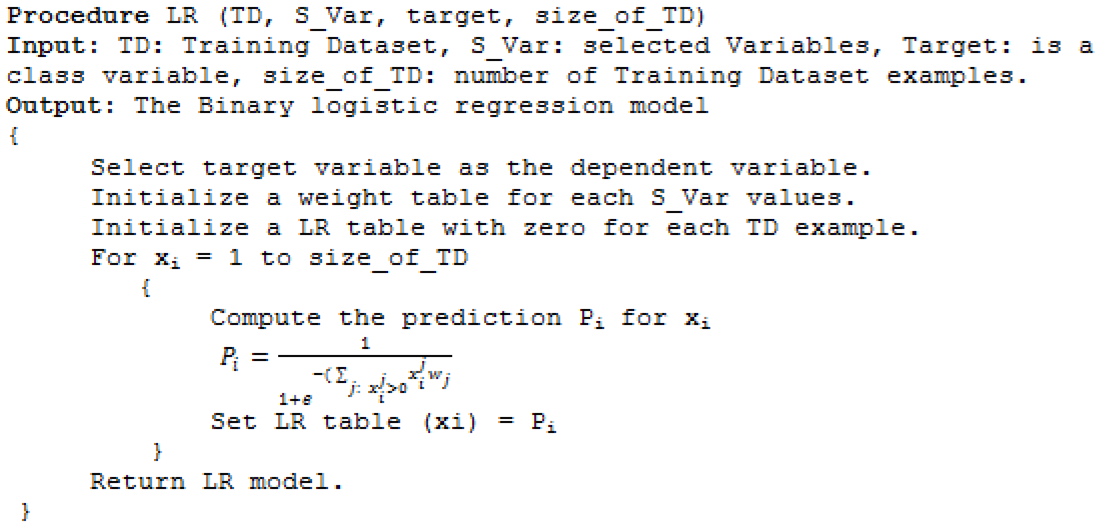
Figure 26: The SVM-RBF algorithm Pseudocode.

3.2.3.4.3. Logistic Regression Model (LR)

The LR model is built as binary classification model for pipes break prediction by rattle package in R language, which used the “glm” package in R. The gml function is a the package available in the R language used to build a many type of Regression models including the Logistic regression model that is that is useful for predicting of a binary dependent variable from a set of continuous predictor variables.

It is used to generalize linear model that is useful to give a description of the error distribution and a symbolic description of the predictor. (Hothorn, 2011).

The LR algorithm Pseudocode is display in Figure (27)



```
Procedure LR (TD, S_Var, target, size_of_TD)
Input: TD: Training Dataset, S_Var: selected Variables, Target: is a
class variable, size_of_TD: number of Training Dataset examples.
Output: The Binary logistic regression model
{
      Select target variable as the dependent variable.
      Initialize a weight table for each S_Var values.
      Initialize a LR table with zero for each TD example.
      For $x_i$ = 1 to size_of_TD
          {
               Compute the prediction $P_i$ for $x_i$
               $P_i = \frac{1}{1+e^{-(\Sigma_{j:\ x_i^j>0} x_i^j w_j}}$
               Set LR table (xi) = $P_i$
           }
      Return LR model.
 }
```

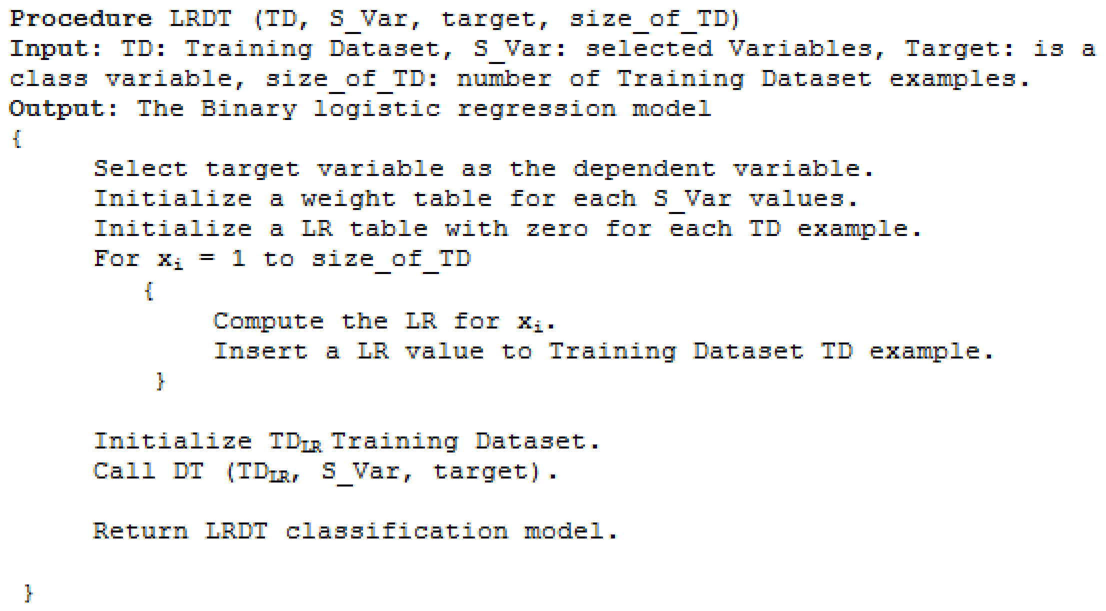
Figure 27: The LR algorithm Pseudocode.

3.2.3.4.4. Logistic Regression Decision Tree model (LRDT)

The LRDT model is the model was built by combined the LR and DT algorithms. The “glm” and 'rpart' packages are used to build this model, with the same parameters that are used with the LR and DT models.

The LRDT model is implemented by executed the logistic regression function on training dataset example then classification data by Decision Tree algorithm.

The LRDT algorithm Pseudocode is display in Figure (28)



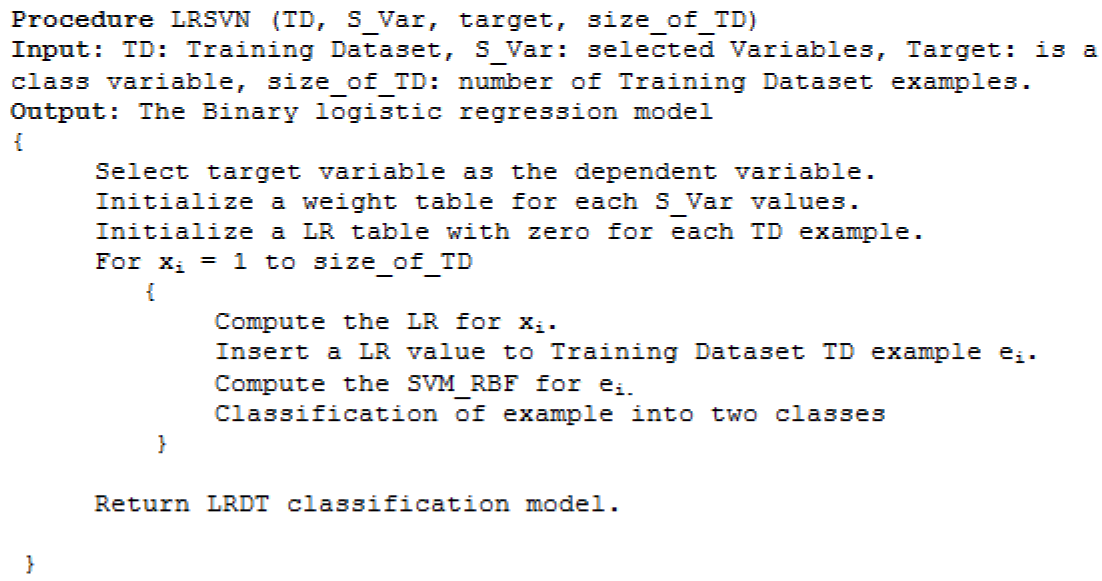
```
Procedure LRDT (TD, S_Var, target, size_of_TD)
Input: TD: Training Dataset, S_Var: selected Variables, Target: is a
class variable, size_of_TD: number of Training Dataset examples.
Output: The Binary logistic regression model
{
      Select target variable as the dependent variable.
      Initialize a weight table for each S_Var values.
      Initialize a LR table with zero for each TD example.
      For $x_i$ = 1 to size_of_TD
          {
               Compute the LR for $x_i$.
               Insert a LR value to Training Dataset TD example.
           }

      Initialize $TD_{LR}$ Training Dataset.
      Call DT ($TD_{LR}$, S_Var, target).

      Return LRDT classification model.

 }
```

Figure 28: The LRDT algorithm Pseudocode.

3.2.3.4.5. Logistic Regression Support Vector Machine Model (LRSVM)

The LRSVM model is the model was built by combined the LR and SVM algorithms. The “glm” and ' kernlab ' packages are used to build this model, with the same parameters that are used with the LR and SVM models.

This model is implemented by executed the logist regression function on training dataset example then classification data by SVM_RBF algorithm. The LRSVM algorithm Pseudocode is display in Figure (29).



```
Procedure LRSVN (TD, S_Var, target, size_of_TD)
Input: TD: Training Dataset, S_Var: selected Variables, Target: is a
class variable, size_of_TD: number of Training Dataset examples.
Output: The Binary logistic regression model
{
     Select target variable as the dependent variable.
     Initialize a weight table for each S_Var values.
     Initialize a LR table with zero for each TD example.
     For $x_i$ = 1 to size_of_TD
        {
             Compute the LR for $x_i$.
             Insert a LR value to Training Dataset TD example $e_i$.
             Compute the SVM_RBF for $e_i$.
             Classification of example into two classes
         }

     Return LRDT classification model.

 }
```

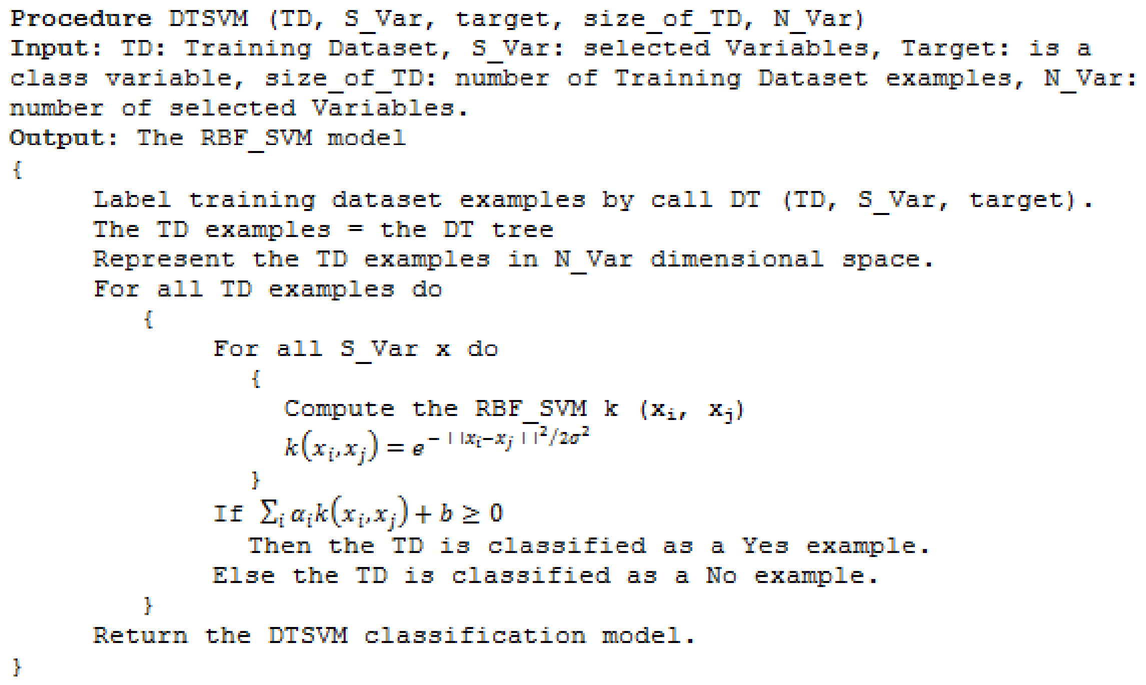
Figure 29: The LRSVM algorithm Pseudocode.

3.2.3.4.6. Decision Tree Support Vector Machine Model (DTSVM)

The DTSVM model is the model was built by combined the DT and SVM algorithms. The “rpart” and ' kernlab ' packages are used to build this model, with the same parameters that are used with the DT and SVM models.

This model is implemented by executed the Decision Tree algorithm on training dataset example then classification data by SVM_RBF algorithm.

The DTSVM algorithm Pseudocode is display in figure (30).



```
Procedure DTSVM (TD, S_Var, target, size_of_TD, N_Var)
Input: TD: Training Dataset, S_Var: selected Variables, Target: is a
class variable, size_of_TD: number of Training Dataset examples, N_Var:
number of selected Variables.
Output: The RBF_SVM model
{
     Label training dataset examples by call DT (TD, S_Var, target).
     The TD examples = the DT tree
     Represent the TD examples in N_Var dimensional space.
     For all TD examples do
         {
             For all S_Var x do
               {
                 Compute the RBF_SVM k ($x_i$, $x_j$)
                 $k(x_i, x_j) = e^{-\,||x_i - x_j||^2/2\sigma^2}$
               }
             If $\sum_i \alpha_i k(x_i, x_j) + b \geq 0$
               Then the TD is classified as a Yes example.
             Else the TD is classified as a No example.
         }
     Return the DTSVM classification model.
}
```

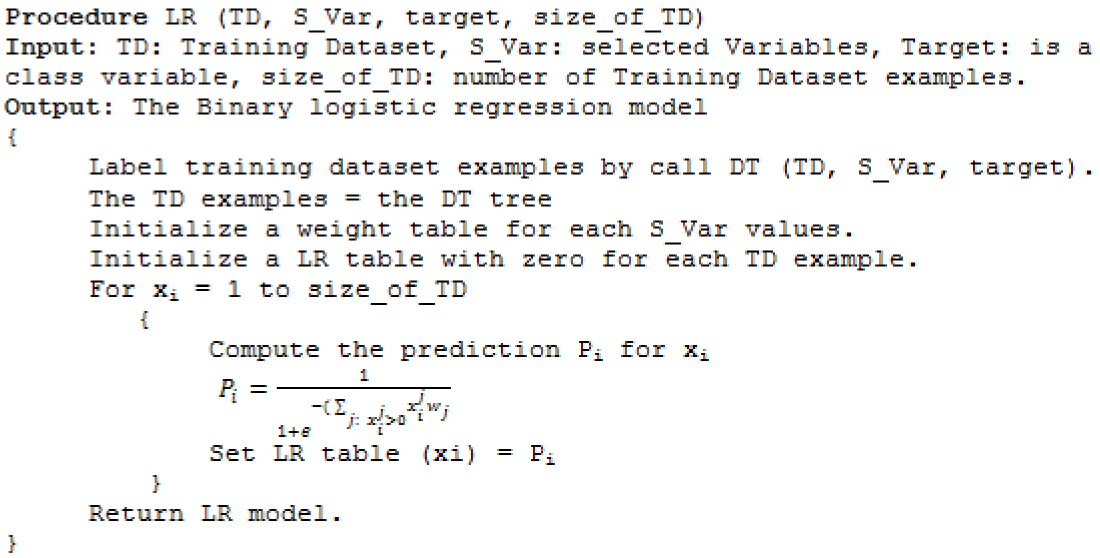
Figure 30: The DTSVM algorithm Pseudocode.

3.2.3.4.7. Decision Tree Logistic Regression Model (DTLR)

The DTLR model is the model was built by combined the DT and LR algorithms. The “rpart” and ”glm” packages are used to build this model, with the same parameters that are used with the DT and LR models.

This model is implemented by executed the Decision Tree algorithm on training dataset example then classification data by LR algorithm.

The DTLR algorithm Pseudocode is display in figure (31).



```
Procedure LR (TD, S_Var, target, size_of_TD)
Input: TD: Training Dataset, S_Var: selected Variables, Target: is a
class variable, size_of_TD: number of Training Dataset examples.
Output: The Binary logistic regression model
{
      Label training dataset examples by call DT (TD, S_Var, target).
      The TD examples = the DT tree
      Initialize a weight table for each S_Var values.
      Initialize a LR table with zero for each TD example.
      For $x_i$ = 1 to size_of_TD
          {
               Compute the prediction $P_i$ for $x_i$
               $P_i = \frac{1}{1+e^{-(\sum_{j:\, x_i^j>0} x_i^j w_j}}$
               Set LR table (xi) = $P_i$
           }
      Return LR model.
}
```

Figure 31: The DTLR algorithm Pseudocode.

Chapter Four Results and Discussion

4.1 Introduction

Seven binary classification models were obtained to predict the break of the pipes based, after on three techniques are the Decision Tree, Logistic Regression and the Support Vector Machine algorithms,

This chapter presents the results obtained from building each one of these models. Then, it displays the Models performance by calculating the Error Rate, Sensitivity, False Positive Rate, Specificity, Precision and Accuracy for each one of these models depending on the Confusion Matrixes obtained from building each model.

The Models performance compared to know the best one, which this research will recommend to use and apply by the authorities concerned with this research.

4.2 Results of Models

In this section the models results are introduced, to show impact of input variables in pipes breaking.

4.2.1 Results of Decision Tree model

When the DT model was built, the results of this model were collected. The model shows that four variables from all the input variables are actually used to construct the DT, which means that they affect pipes breaking, which are pressure zone of pipe, age, material and diameter.

The DT structure was also obtained for the decision tree model, which displays the cases when the pipe will break or not with the probability of it, depending on the values of the four variables that affect pipes and may cause breaking for them.

The summary of rules that give us possibility for pipes breaking are:

1. If Pressure_Zone =Asker_Camps,Ein_Camp,Rafedia,S5 or Zawata, Diam>=95 and pipe.age>=35 then the probability of pipe break = 0.95

2. If Pressure_Z=C1, E0.1, E0.2, E0.3, Gasil, Hurush_line, NE1, NE2, NE4, NW0, NW1+NW2a, NW3+NW2b, out nablus, S2, S3, S4, SE1, SE2, SE3, W1, W0, W1, W2a, W2b, W3, W4, Worash_Gravity and pipe.age< 49.5 and Material=DI,GS,HDPE,PVC or if Pressure_Z=C1,E0.2,NE1,S3,S4,W-1,W0,W1,W2a,W3,W4 and Type=dis_2 and Material = HDPE then the probability of pipe break = 0.91.

3. If Pressure_Z=Asker_Camps,Ein_Camp,Rafedia,S5,Zawata and Diam< 95 and pipe.age>=11.5 then the probability of pipe break = 0.91.

4. If Pressure_Z=Asker_Camps,Ein_Camp,Rafedia,S5,Zawata and Diam< 9 and pipe.age< 11.5 and Material=HDPE then the probability of pipe break = 0.89

5. If Pressure_Z=C1 , E0.1 , E0.2, E0.3, Gasil, Hurush_line, NE1, NE2, NE4, NW0, NW1+NW2a, NW3+NW2b, out nablus, S2, S3, S4, SE1, SE2, SE3, W-1, W0, W1, W2a, W2b, W3, W4, Worash_Gravity and pipe.age< 49.5 and Material=PP,ST and pipe.age>=12.5 then the probability of pipe break = 0.85

To understand the details of the decision tree structure and all the rules that represent this tree, see the Appendix A

4.2.2 Results of Support Vector Machine (SVM) model

After the SVM model was built the results were obtained show that the characteristics of this model are number of the support vectors (i.e., the number of observations are on the boundary), the value of the objective function (which is the algorithm optimizes) and the error value, that is calculated on the training dataset. The SVM model turns out that the it contains 487 of Support Vectors, -294.9771 objective function values and 0.031889 error value on the training dataset.

4.2.3 Results of Logistic Regression (LR) Model

The Logistic Regression model results show that the factors which have a significant impact on the occurrence of broken pipes in the water distribution network, which ranked by importance are Pressure Zone values, type of network, pipe age, Material and pipe diameter.

The pressure zones are arranged according to their effects as follows: out Nablus, SE3, Hurush_line, SE2, SE2, NW0, Worash_Gravity, E0.3, W2b, S2, W2a, E0.1, W1, S3, W0, NE2, E0.2, W4, S5, W-1, SE1, NW1+NW2a, Rafedia, NE1, W3, S4, Zawata, C1, NW3+NW2b and Ein_Camp.

This result shows that if the pipe is in the old water distribution network, the odds of pipe breaks will increase by 5.63. The pipes made from HDPE material increase the odds of pipe breaks by 0.38.

4.2.4 Results of LRDT, LRSVM, DTSVM and DTLR Models

These models show that the variables that affect the broken pipes ranked by importance are: Pressure Zone values, type of network, pipe age, Material and pipe diameter. And it present classification models to predict the broken pipes and many of rules to that,

4.3 Models Performance Evaluation

The evaluation model performance is the next step to build models on data mining. The models that have been built should be validated, examined and evaluated to ensure their accuracy, efficiency and feasibility in achieving the objective of the research.

Models performance is evaluated by obtaining a confusion matrices for all models, then the Error Rate, Sensitivity (True Positive Rate), False Discovery Rate, Miss Rate (False Negative Rate) Precision (Positive Predictive Value), Specificity (True Negative Rate) and Accuracy performance measures are calculated for each model to evaluate and compare the models performance, where the error rate shows a rate of wrong predictions in overall predictions, the Sensitivity is the amount of lost water that the model can protect it, the False discovery rate is the rate of pipes that the system expected they will be broken, but they actually did not. Which represents the maintenance recommended by the system but actually did not need to be conducted at the moment, the Miss Rate represents the amount of lost water from the water distribution network due to broken pipes, which the model could not protect it. The Precision represents the pipes that were actually broken and the model succeeded in detecting them, the Specificity is the number of pipes that have not actually broken, and that the models have successfully predicted them and the Accuracy is the ability of the model to give correct prediction in overall.

The models performance evaluation process in our research is divided into two phases. The first phase is to test the models on the testing data set. The second phase is models testing and validation on maintenance data from 2011 to 2016. The one of the most prominent performance measures is the confusion matrix, which use to calculate many of performance measurements.

In next section the models performance evaluation phases are performed.

4.3.1 The Models Performance Evaluation on Testing Data

After the building of the prediction models on the training data set are completed. The testing data set, which is consisting of 961 records were selected randomly, that is fed into the models and obtained their results. Since these data are known, the prediction results are compared with the actual results in the data set in operation, and this comparison or differentiation leads to verification of the accuracy of the model through different methods of evaluation. In this section, these models were performed on the test data set to evaluate the performance of the all models that were built. Then, the confusion matrix was obtained for each of these models, after that, the models are evaluated by calculating the number of performance measurements.

The confusion matrices for each of the all prediction models on this data show that the testing data set is actually classified into two classes. They are: The "Yes" class, which contains the pipes that were actually broken, that are 187 pipes, and "No" class, which contains the pipes that were actually not broken, that is 774 pipes, the summary of these matrices shows in Table (4) .

Table 4: The Summary for confusion matrices of all models on testing dataset.

		TN	FP	FN	TP
Model	DT	761	13	41	146
	SVM	762	12	24	163
	LR	742	32	38	149
	LRDT	766	14	14	167
	LRSVM	773	7	1	180
	DTSVM	796	6	5	174
	DTLR	794	8	9	150

The confusion matrix of the DT for pipe breaks prediction model shows this model predicted that 159 pipes would be broken while the 802 pipes would not be broken. This classification model succeeded in predicting that 761 pipes would not be broken, and it actually did not break, and 146 pipes would be broken and it actually broke. However, it failed to predict on the 13 pipes. It predicted they would be broken but actually they didn't, it is also failed to predict on 41 pipes when it predicted they would not be broken, but they broke.

The confusion matrix of the SVM model shows this model predicted that 175 pipes would be broken while the 786 pipes would not be broken. It succeeded in predicting that 762 pipes would not be broken, and it actually did not break, and 163 pipes would be broken and it actually broke. However, it failed to predict on the 12 pipes. It predicted they would be broken but actually they didn't, it is also failed to predict on 24 pipes when it predicted they would not be broken, but they broke

The confusion matrix of the LR model shows this model predicted that 181 pipes would be broken while the 780 pipes would not be broken. This classification model succeeded in predicting that 742 pipes would not be broken, and it actually did not break, and 149 pipes would be broken and it actually broke. However, it failed to predict on the 32 pipes. It predicted they would be broken but actually they didn't, it is also failed to predict on 38 pipes when it predicted they would not be broken, but they broke.

Table The LRDT model predicted that 181 pipes would be broken while the 780 pipe would not break. This classification model succeeded in predicting that 766 pipes would not be broken, and it actually did not break, and 167 pipes would be broken and it actually broke. However, it failed to predict on the 14 pipes. It predicted they would be broken but actually they didn’t, it is also failed to predict on 14 pipes when it predicted they would not be broken, but they broke.

The confusion matrix for the LRSVM model succeeded in predicting that 773 pipes would not be broken, and it actually did not break, and 180 pipes would be broken and it actually broke. However, it failed to predict on the 7 pipes. It predicted they would be broken but actually they didn’t, it is also failed to predict on 1 pipe when it predicted they would not be broken, but they broke.

The confusion matrix for the DTSVM model succeeded in predicting that 796 pipes would not be broken, and it actually did not break, and 174 pipes would be broken and it actually broke. However, it failed to predict on the 6 pipes.

It predicted they would be broken but actually they didn’t, it is also failed to predict on 5 pipes when it predicted they would not be broken, but they broke.

The confusion matrix for the DTSVM model shown That succeeded in predicting that 794 pipes would not be broken, and it actually did not break, and 150 pipes would be broken and it actually broke.

However, it failed to predict on the 8 pipes. It predicted they would be broken but actually they didn’t, it is also failed to predict on 9 pipes when it predicted they would not be broken, but they broke.

The result of all performance measures for models on testing dataset shown in the Table (5).

Table 5: Models performance on testing dataset.

		Model						
		DT	LR	SVM	LRDT	LRSVM	DTSVM	DTLR
Measure	Error rate	0.06	0.07	0.04	0.03	0.008	0.011	0.018
	Sensitivity	0.78	0.80	0.87	0.92	0.99	0.97	0.94
	false discovery rate	0.08	0.18	0.06	0.08	0.03	0.03	0.05
	Miss rate	0.22	0.20	0.13	0.08	0.01	0.03	0.06
	Precision	0.92	0.82	0.93	0.92	0.96	0.96	0.95
	Specificity	0.98	0.96	0.98	0.98	0.99	0.99	0.99
	Accuracy	0.94	0.93	0.96	0.97	0.99	0.98	0.98

which shows that the error rate for the LRSVM model is the lowest compared to the error rate of the other models in the Table ,since its error rate is negligible, The Sensitivity metric shows that the LRSVM model is the best one of these models, where it provides approximately 99% from lost water from broken pipes, the False Discovery Rate shows that the LRSVM and DTSVM models have the same false positive rate value that is the least for all these models, the Miss Rate shows that the LRSVM model has the lowest value for all these models, Best specificity value in Table is 0.99, where it was achieved by the LRSVM, DTSVM and DTLR models, The LRSVM model the most accurate of these models. As a final result for using all previous performance measures to evaluate the performance of all the building models, the LRSVM model is the best model for pipes breaks prediction compared with other models in this research.

4.3.2 The Models Performance Evaluation on All Available Maintenance Data

All prediction models for broken pipes that were built in this research; especially, the LRSVM model that is proved to be highly efficient when examined on a test data set. However, this is not enough to judge the efficiency of this model as the size of the training data set is rather small. As a result, these models must be applied on a large and real data set, and then evaluate the performance of the models on it; to know their efficiency in fact.

Therefore, this section assesses the performance of most of these models on maintenance data for the years 2011 to 2016. Through building pipe break prediction models using data for years before 2011, and then evaluating model performance through the use of maintenance data set for year 2011 as a test data set. The models are then built by adding the maintenance data set for year 2011 to the models, and then tested on the 2012 data. After that, all maintenance data sets until 2012 are used to build the models, and then use the maintenance data set for 2013 as a testing data set to evaluate models performance. Then do the same way applied with all the data in the following years until 2016. The overview of all data sets testing is shown in Figure (32).

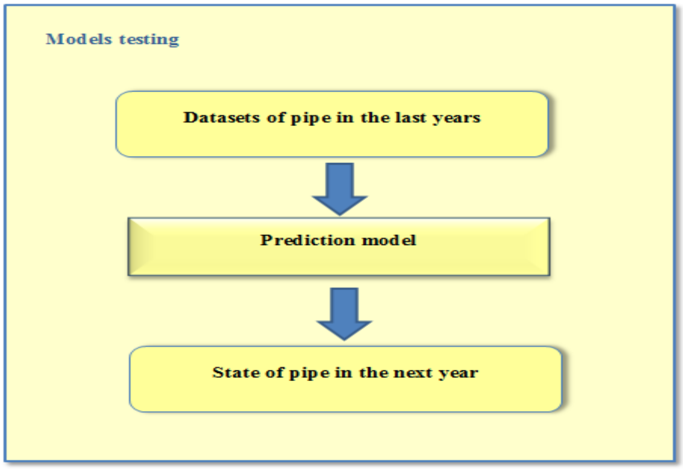

Models testing
Datasets of pipe in the last years
Prediction model
State of pipe in the next year

Figure 32: Model testing overview.

The result of confusion matrices for DT, SVM, LR, LRSVM models on maintenance data for the years from 2011 to 2016 shown in Table (6).

Table 6: The result of confusion matrices for models on maintenance data.

Dataset	Model	TN	FP	FN	TP
2011	DT	3442	379	210	301
	LR	3346	457	118	393
	SVM	3421	400	109	402
	LRSVM	3431	479	43	379
2012	DT	3902	113	310	280
	LR	3902	113	281	309
	SVM	3855	160	165	425
	LRSVM	3950	15	235	405
2013	DT	3612	71	250	676
	LR	3516	167	355	571
	SVM	3632	51	323	603
	LRSVM	3585	0	382	642
2014	DT	3743	102	338	608
	LR	3632	213	244	702
	SVM	3741	104	209	737
	LRSVM	3758	33	22	978
2015	DT	3649	97	238	812
	LR	3536	210	200	850
	SVM	3652	94	180	870
	LRSVM	3677	27	121	971
2016	DT	3670	113	216	808
	LR	3543	240	190	834
	SVM	3682	101	186	838
	LRSVM	3681	128	28	970

The same performance measures that were calculated when the models were executed on the testing data set are computed at this stage, depending on the all values of confusion matrices for models. The DT, LR and SVM models performance on all datasets show in Table (7), and the SVM and LRSVM models performance on all datasets show in Table (8)

Table 7: The performance of DT. LR and SVM models on data sets for the years 2011 to 2016.

		Model																	
		2011			2012			2013			2014			2015			2016		
		DT	SVM	LR	DT	SVM	LR	DT	SVM	LR	DT	SVM	LR	DT	SVM	LR	DT	SVM	LR
Measure	Error rate	0.14	0.12	0.13	0.09	0.07	0.09	0.07	0.08	0.11	0.09	0.07	0.10	0.07	0.06	0.09	0.07	0.06	0.09
	Sensitivity	0.59	0.79	0.77	0.47	0.72	0.52	0.73	0.65	0.62	0.64	0.78	0.74	0.77	0.83	0.81	0.79	0.82	0.81
	false discovery rate	0.56	0.50	0.54	0.29	0.27	0.27	0.10	0.08	0.23	0.14	0.12	0.23	0.11	0.10	0.20	0.12	0.11	0.22
	Miss rate	0.41	0.21	0.23	0.53	0.28	0.48	0.27	0.35	0.38	0.36	0.22	0.26	0.23	0.17	0.19	0.21	0.18	0.19
	Precision	0.44	0.50	0.46	0.71	0.73	0.73	0.90	0.92	0.77	0.86	0.88	0.77	0.89	0.90	0.80	0.88	0.89	0.78
	Specificity	0.90	0.90	0.88	0.97	0.96	0.97	0.98	0.99	0.95	0.97	0.97	0.94	0.96	0.97	0.94	0.97	0.97	0.94
	Accuracy	0.86	0.88	0.87	0.91	0.93	0.91	0.93	0.92	0.89	0.91	0.93	0.90	0.93	0.94	0.91	0.93	0.94	0.91

Table 8: The performance of SVM and LRSVM models on data sets for the years 2011 to 2016.

		Model											
		2011		2012		2013		2014		2015		2016	
		SVM	LRSVM	SVM	LRSVM	SVM	LRSVM	SVM	LRSVM	SVM	LRSVM	SVM	LRSVM
Measure	Error rate	0.12	0.12	0.07	0.05	0.08	0.08	0.07	0.01	0.06	0.03	0.06	0.03
	Sensitivity	0.79	0.90	0.72	0.63	0.65	0.63	0.78	0.98	0.83	0.89	0.82	0.97
	false discovery rate	0.50	0.56	0.27	0.04	0.08	0.00	0.12	0.03	0.10	0.03	0.11	0.12
	Miss rate	0.21	0.10	0.28	0.37	0.35	0.37	0.22	0.02	0.17	0.11	0.18	0.03
	Precision	0.50	0.44	0.73	0.96	0.92	1.00	0.88	0.97	0.90	0.97	0.89	0.88
	Specificity	0.90	0.88	0.96	1.00	0.99	1.00	0.97	0.99	0.97	0.99	0.97	0.97
	Accuracy	0.88	0.88	0.93	0.95	0.92	0.92	0.93	0.99	0.94	0.97	0.94	0.97

Depending on the previous two Tables, the Error Rate for each model for all study data sets shows the SVM model has the lowest error rate for most years compared with other models in Table (7) , since the error rate for it does not exceed 12% for all data sets. But the LRSVM model has the low or the same error rate for SVM model. the Sensitivity show that the SVM model is the better than DT and LR models on all data sets except for the 2013, because there is have a missing data, where the SVM model provides approximately 72% to 83% from lost water. when compared the SVM and LRSVM Sensitivity show that the LRSVM model Sensitivity is better than SVM in most datasets that have been tested. The SVM model has lowest False discovery rate compared with DT and LR models, but LRSVM has a false discovery rate of less than the SVM model. the SVM model have the lowest miss rate for the three models on all datasets except for the 2013 dataset because there is have a missing data. Where the SVM model have approximately 17% to 28% from miss rate, which is a small value compared to the amount of water lost that protects it, the LRSVM model have the almost low miss rate than SVM for all datasets. The SVM model has the highest Precision compared to the other two models and the LRSVM Precision is better than SVM Precision.

The SVM model has the highest Specificity compared to the other two models on all data sets exception on the data for 2012, but the difference between it and the LR model is 0.01 that is a small difference. The LRSVM Specificity is better than SVM Specificity. The SVM model has the accuracy better than other two models on all datasets except on the 2013 dataset because there is have a missing data. The LRSVM model has the best accuracy for all these models, its accuracy is good.

The summary of models performance evaluation is the LRSVM model is the best one of these models that can be used for pipes break prediction.

Chapter Five

Conclusion and Future Work

5.1 Introduction

In this research, data mining techniques were applied to the data which was obtained from the Studies and Maintenance departments from Nablus Municipality in order to develop a model for the prediction of pipe break by testing, and then later classifying the pipe patterns in the water distribution network in Nablus city into two categories, namely pipes with break, and pipes without break. The importance of this research in addition to the recommendations and the future work are presented in the following sections.

5.2 Contributions and Significance of the Research

- ✓ The performance evaluation of research models shows that they are effective for predicting pipe breakage because they resulted in high accuracy with a little error rate in prediction in addition to their ability to save a good amount of water that is lost as a result of broken pipes in the network .
- ✓ When comparing the performance of research models, it was found that the LRSVM model is the most reliable model in the anticipation of pipes' breaks because it gave the best values for most of the calculated performance measures as its error rate varied between 0.01 and 0.12, and it may be able to save up to 0.97 water from the amount of water lost from the network, with an accuracy rate that may reach 0.99.
- ✓ This research presented an integrated database for the water distribution network in Nablus.
- ✓ The research models presented the variables that affect the broken pipes, which are Pressure Zone values, type of network and pipe age, material and diameter.

5.3 Recommendations

Based on the results of this research, there is a number of recommendations to be submitted to Nablus Municipality in order to effectively solve the problem of water loss from the water distribution network. The recommendations are as follows:

1. Adopting the integrated database, this was built during this research by the Department of Water and Sanitation in the Municipality of Nablus in both Maintenance and Studies sections.
2. Employing the predictive models in order to allocate the budget of the maintenance department. This can be realized by determining the number of pipes that are expected to break during a certain period of time and the specifications of these pipes, thus helping in the estimation of maintenance requirements for the broken pipes in the water distribution network.
3. Utilizing prediction models to minimize the problem of water loss from broken pipes, through carrying out maintenance work on the pipelines that are expected to break, before they break.
4. Designing an electronic form for the old water distribution network in order to retrieve missing pipes data.
5. Collecting data about the pipes underground depth and adding them to the records of the Water and Sanitation Department in order to use it to study the impact of this factor on the pipes break.

6. Determining by installing pressure sensors on the network, the pressure value in each pipe in the network accurately, rather than relying on the pressure zone in general, as the pressure of water in each pipe in any area varies between the pipes and as the pressure at the beginning of the pumping area is higher than it at the end.
7. Benefiting from the Global Positioning System (GPS) through the use of an application that would help determine the coordinates of the piping site under maintenance and connect the maintenance work with the Geographic Information Systems (GIS) network.
8. Designing an online system that relies on this research's integrated database to connect the maintenance and studies departments' data with the maintenance staff in different work zones directly.

5.4 Future Work

Although this research presented a good model for predicting pipe breakage, this does not mean that research stops here; on the contrary, this research opens up prospects for future works as shown below:

- ✓ Research can be further expanded by collecting more longitudinal data on pipes breakage and by adding the parameters of pipes depth and pressure value to increase the accuracy of prediction performance models.

- ✓ This study focused on the study and analysis of data on the main water pipelines in the water distribution network in Nablus City. Future research can be conducted to study all the pipes in the water distribution network.

- ✓ The results of this research can be used in the future to build an application that would help in the prediction of pipe breaking and that would assist the maintenance department in budgeting and assessing the maintenance equipment for use in the Maintenance and Studies departments in Nablus Municipality.

- ✓ The GIS system can be used to represent the water distribution network which is implemented by Nablus Municipality in order to use the spatial data mining techniques to build a model for predicting pipe breakage in the water distribution network. The model which will utilize the GPS system should enable its users to locate the pipes that are expected to break. This would in turn facilitate and speed up the work of the maintenance teams, and reduce the problem of water loss from broken pipes in the water distribution network.

References

Bubtiena, A. M., ElShafie, A. H., & Jaafar, O. (2011, October). Performance improvement for pipe breakage prediction modeling using regression method. International Journal of the Physical Sciences, 6(25), 6025-6035.

Francis, R. A., Guikema, S. D., & Henneman, L. (2014). Bayesian Belief Networks for predicting drinking water distribution system pipe breaks. *Reliability Engineering & System Safety, 130*, 1–11.

Hart, P. D. (2016). Receiver Operating Characteristic (ROC) Curve Analysis: A Tutorial Using Body Mass Index (BMI) as a Measure of Obesity. *Journal of Physical Activity Research, 1*(1), 5-8.

Karatzoglou, A., Meyer, D., & Horni, K. (2006, April). Support Vector Machines in R. *Journal of Statistical Software, 15*(9).

Martins, A. C., Leitão, J. P., & Amado, C. (2011, December). Comparative Study of Three Stochastic Models for Prediction of Pipe Failures in Water Supply Systems. *Journal of Infrastructure Systems, 19*(4), 442–450.

Martins, A. C., Leitão, J. P., & Amado, C. (2013, December). Comparative Study of Three Stochastic Models for Prediction of Pipe Failures in Water Supply Systems. *Journal of Infrastructure Systems, 19*(4), 442–450.

Prajwala, T. R. (2015, January). A Comparative Study on Decision Tree and Random Forest Using R Tool. *International Journal of Advanced Research in Computer and Communication Engineering, 4*(1).

Sokolova , M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management, 45*, 427–437.

Venables, W. N., & Smith, D. M. (2017). *An Introduction to R.* Notes on R: A Programming Environment for Data Analysis and Graphics, for R, version 3.4.1 , R Core Team.

Wang, R., Dong, W., Wang, Y., Tang, K., & Yao, X. (2013, June). Pipe failure prediction: A data mining method. *International Conference on Data Engineering (ICDE)*.

Xu, Q., Chen, Q., Li , W., & Ma, J. (2011, January). Pipe break prediction based on evolutionary data-driven methods with brief recorded data. *Reliability Engineering & System Safety, 96*(8), 942–948.

(1998). *Support Vector Machines for Classification and Regression.* Technical Report, University Of Southampton.

(2012, Augast). Retrieved from kdnuggets: http://www.kdnuggets.com/2012/08/poll-analytics-data-mining-programming-languages.html

Field Data–Based Methodology for Estimating the Expected Pipe Break Rates of Water Distribution Systems. (2016). *Journal of Water Resources Planning and Management, 142*(10).

Field Data–Based Methodology for Estimating the Expected Pipe Break Rates of Water Distribution Systems. (2016, Octoper). *Journal of Water Resources Planning and Management, 142*(10).

Adel Al-Salaymeh, I. A.-K. (2009). Quality of Drinking Water from Rainwater Harvesting Cisterns of Hebron City and Factors Affecting It. *2nd International Conference on the Palestinian Environment*.

Arasmith, S. (2009). *introduction to small water system.* USA.

Arora, J., Bhalla, N., & Rao, S. (2013, July). A Review on Association Rule Mining Algorithms. *International Journal of Innovative Research in Computer and Communication Engineering, 1*(5).

Bhavsar, H., & Panchal, M. H. (2012, December). A Review on Support Vector Machine for Data Classification. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 1*(10).

Bordley, R. F. (2002, May). Decision Rings: Making Decision Trees Visual and non-Mathematical. *INFORMS Transactions on Education, 2*(3).

Boser, B. E., Guyon, I., & Vapnik, V. N. (1992). A Training Algorithm for Optimal Margin Classifier. *5th Annual ACM Workshop on COLT*, 144-152.

Chakrabarti, S., Ester, M., Fayyad, U., Gehrke, J., Han, J., Morishita, S., et al. (2006, April). Data Mining Curriculum: A Proposal (Version 1.0). *ACM SIGKDD*.

(2011). Random Forests. In A. Cutler, D. R. Cutler, J. R. Stevens, C. Zhang, & Y. Ma (Eds.), *Ensemble Machine Learning: Methods and Applications* (pp. 157-176). USA, New York: Springer.

Das, S. P., & Padhy, S. (2012, March). Support Vector Machines for Prediction of Futures Prices in Indian Stock Market. *International Journal of Computer Applications, 41*(3).

David G. Kleinbaum, M. K. (2010). *Logistic Regression A Self-Learning Text.* Atlanta: Springer.

Dong , G., & Pei, J. (2007). *Sequence Data Mining* (1 ed.). USA, New York: Springer Science+Business Media.

Esling, P., & Agon, C. (2012, November). Time-Series Data Mining. *ACM Computing Surveys*.

Fawcett, T. (2005, December). An introduction to ROC analysis. *Pattern Recognition Letters*, 861–874.

Ghorbanian, V., Guo, Y., & Karney, B. (2016, Octoper). Field Data–Based Methodology for Estimating the Expected Pipe Break Rates of Water Distribution Systems. *Journal of Water Resources Planning and Management, 142*(10).

Giustolisi, O., Savic, D. A., & Laucelli, D. (2004). Data Mining for Management and Rehabilitation of Water Systems: The Evolutionary Polynomial Regression Approach. *Dresdner Wasserbauliche Mitteilungen.*

Gupta, M., & Aggarwal, N. (2010, March). Classification Techniques Analysis. *National Conference on Computational Instrumentation*.

Halili, F., & Rustemi, A. (2016, August). Predictive Modeling: Data Mining Regression Technique Applied in a Prototype. *International Journal of Computer Science and Mobile Computing, 5*(8).

Hamalainen, W. (2006). *Descriptive and Predictive Modelling Techniques for Educational.* Licentiate thesis, Joensuu University, Joensuu, Finland.

Han, J., & kamber, M. (2012). *data mining: concept and techniques.* (3, Ed.) USA, San Francisco, CA: Elsevier Publishers inc.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining (Adaptive Computation and Machine Learning)* (1 ed.). A Bradford Book.

Heeger, D. (1997, November). Signal Detection Theory.

Hothorn, B. S. (2011). *A Handbook of Statistical Analyses.* Retrieved 11 5, 2017, from cran.r-project: https://cran.r-project.org/web/packages/HSAUR/vignettes/Ch_logistic_regression_glm.pdf

Jabari, S. J. (2017). Non-Revenue Water Management in Palestine. *International Journal of Urban and Civil Engineering*, Vol:11, No:7.

Jad Isaac, K. R. (2015). status of the environment in the state of palestine. *the Applied Research Institute – Jerusalem (ARIJ)*.

Kalechofsky, H. (September, 2016). *A Simple Framework for Building Predictive Models* (1 ed.). A Solomon Edwards.

Kapoor, A. (2014, February). Data Mining: Past, Present and Future Scenario. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 3*(1).

Kumar, P., Krishna, P. R., & Raju, S. B. (2012). *Pattern Discovery Using Sequence Data Mining: Applications and Studies* (1 ed.). USA, Hershey: Information Science Reference (an imprint of IGI Global).

Liaw, A., & Wiener, M. (2002, December). Classification and Regression by randomForest. *R News, 2/3*.

Lorenz, J. R. (2015). *Modeling Binary Correlated Responses using SAS, SPSS and R .* Switzerland: Springer International Publishing.

Louppe, G. (2014). *Understanding Random Forest from Theory to Practice.* PhD Thesis, University of Liège, Liège.

Mashford, J., Silva, D. D., Marney, D., & Burn, S. (2009, November). An approach to leak detection in pipe networks using analysis of monitored pressure values by support vector machine. *Third International Conference on Network and System Security*, 534-539.

Mayoraz, E., & Alpaydın, E. (1999). Support Vector Machines for Multi-class Classification. *International Work-Conference on Artificial and Natural Neural Networks.* Alicante, Spain.

Mielikainen, T. (2005). *Summarization Techniques for Pattern Collections in Data Mining.* Academic Dissertation, Helsinki University, Helsinki.

Mlambo, N. (2016, March). Data Mining: Techniques, Key Challenges and Approaches for Improvement. *International Journal of Advanced Research in, 6*(3).

Neelamegam, S., & Ramaraj, E. (2013, September). Classification Algorithm in Data mining: An Overview. *International Journal of P2P Network Trends and Technology (IJPTT), 4*(8).

Ng, A. (2016). Retrieved from coursera: https://www.coursera.org/learn/machine-learning

Park, H. S., Jo, C.-H., & Goo, J. M. (2004, February). Receiver Operating Characteristic (ROC) Curve: Practical Review for Radiologists. *Korean Journal of Radiol, 5*(1), 11-18.

Park, H.-A. (2013). An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain. *J Korean Acad Nurs*, 154-164.

Piatetsky, G. (2014, Augaste). Retrieved from KDnuggets: http://www.kdnuggets.com/2014/08/four-main-languages-analytics-data-mining-data-science.html

Ramageri, B. M. (2010, December). Data Mining Techniques And Application. *Indian Journal of Computer Science and Engineering, 1*(4).

Rattrout, A., & Saleh, M. (2016). Prediction of pipe breaks in water distribution system using data mining tools. *INTIS conference.* Moroco conference.

Rokach, L., & Maimon, O. (2014). *Data Mining With Decision Trees Theory and Applications* (2 ed.). UK, Covent Garden, London: World Scientific Publishing Co. Pte. Ltd.

Sondwale, P. P. (2015, April). Overview of Predictive and Descriptive Data Mining Techniques. *International Journal of Advanced Research in Computer Science and Software Engineering, 5*(4).

Strobl, C., Malley, J., & Tutz, G. (2009). *An Introduction to Recursive Partitioning.* Technical Report Number 55, University of Munich, Munich, Germany.

Syed, A. R. (2011). *A Review of Cross Validation and Adaptive Model Selection.* Mathematics Theses, Georgia State University, Georgia, USA.

Tamilselvi, R., & Kalaiselvi, S. (2013, February). An Overview of Data Mining Techniques and Applications. *International Journal of Science and Research (IJSR), 1*(2).

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining* (1 ed.). Indiana, Indianapolis: Pearson Addison-Wesley.

Team, R. C. (2016, Septemper). Retrieved from R: https://cran.r-project.org/

Therneau, T. M., Atkinson, E. J., & Foundation, M. (2017, March). An Introduction to Recursive Partitioning. *R News*.

UNESCO. (2015). *Water for a sustainable world.* Paris: The United Nations World Water Development.

Velickov, S., & Solomatine, D. (2000). Predictive Data Mining: Practical Examples. *Artificial Intelligence in Civil Engineering Proc. 2nd Joint.* Cottbus, Germany.

Walsh, S. (2016). Binary Logistic Regression – What, When, and How. *JMP Discovery Conference.* Cary: SAS Institute.

Wang, F. (2014). *Nutrition Screening Tools and the Prediction of Clinical Outcomes Among Chinese Hospitalized Gastrointestinal Disease Patients.* SURFACE Theses, Syracuse University, New York.

Wang, L. (2005). *Support Vector Machines: Theory and Applications* (1 ed.). Berlin: Springer.

wikipedia. (2017, December 7). *Logistic regression*. Retrieved December 9, 2017, from wikipedia: https://en.wikipedia.org/wiki/Logistic_regression

Williams, G. J. (2009, December). Rattle: A Data Mining GUI for R. *R Journal, 1*(2).

Witten, I., & Frank, E. (2005). *Data mining: Practical Machine Learning Tools And Techniques* (2 ed.). USA, San Francisco, CA: Morgan Kaufmann Inc.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., et al. (2008, January). Top 10 algorithms in data mining. *Knowledge and Information Systems, 14*(1), 1–37.

Xu, Q., Chen, Q., Ma, J., & Blanckaert, K. (2013, March). Optimal pipe replacement strategy based on break rate prediction through genetic programming for water distribution network. *Journal of Hydro-environment Research, 7*(2), 134–140.

Zhao, Y. (2011, April). *R and Data Mining: Examples and Case Studies* (1 ed.). Elsevier.

Zhao, Y. (2012). *R and Data Mining: Examples and Case Studies* (1 ed.). China: Academic Press, Elsevier.

Ziad Mimi, O. A. (2004). Evaluation of Water Losses in Distribution Networks: Rammallah as a Case Study. *Water Science and Technology: Water Supply* , Vol 4 No 3 pp 183–195.

Appendices

A.Experimental results for decision tree model

A.1 The structure of the decision tree

The summary of the structure of the decision tree was obtained after build a classification decision tree model for this research as presented in the next:

```
Summary of the Decision Tree model for Classification (built using 'rpart'):

n= 2885



node), split, n, loss, yval, (yprob)

      * denotes terminal node



 1) root 2885 532 no (0.815597920 0.184402080)

   2) Pressure_Z= C1, E0.1, E0.2, E0.3, Gasil, Hurush_line, NE1, NE2, NE4, NW0,
NW1+NW2a, NW3+NW2b, out nablus, S2, S3, S4, SE1, SE2, SE3, W-1, W0, W1, W2a,
W2b, W3, W4, Worash_Gravity  2280 174 no (0.923684211 0.076315789)

     4) pipe.age< 49.5 2237 132 no (0.940992401 0.059007599)

       8) Material=DI,GS,HDPE,PVC 1965  70 no (0.964376590 0.035623410)

        16) Pressure_Z= E0.1, E0.3, Gasil, Hurush_line, NE2, NE4, NW0, NW1+NW2a,
NW3+NW2b, out nablus, S2, SE1, SE2, SE3, W2b, Worash_Gravity       925    7 no
(0.992432432 0.007567568) *

        17) Pressure_Z= C1, E0.2, NE1, S3, S4, W-1, W0, W1, W2a, W3, W4   1040  63 no
(0.939423077 0.060576923)

          34) Type=dis_1 435   6 no (0.986206897 0.013793103) *
        35) Type=dis_2   605  57   no (0.905785124 0.094214876)

          70) Material=GS, PVC 582  36 no (0.938144330 0.061855670) *

          71) Material=HDPE 23   2 yes (0.086956522 0.913043478) *

      9) Material=PP, ST 272  62 no (0.772058824 0.227941176)

       18) pipe.age< 12.5 205   5 no (0.975609756 0.024390244) *

       19) pipe.age>=12.5 67  10 yes (0.149253731 0.850746269) *

     5) pipe.age>=49.5 43   1 yes (0.023255814 0.976744186) *

   3) Pressure_Z= Asker_Camps, Ein_Camp, Rafedia, S5, Zawata    605   247 yes
(0.408264463 0.591735537)

     6) Diam>=95   208     21   no (0.899038462 0.100961538)

      12) pipe.age< 35   189   3   no (0.984126984 0.015873016) *

      13) pipe.age>=35  19     1    yes (0.052631579 0.947368421) *

     7) Diam< 95 397    60    yes (0.151133501 0.848866499)

      14) pipe.age< 11.5 43  16 no (0.627906977 0.372093023)

        28) Material=GS, PP, ST    25   0   no (1.000000000 0.000000000) *

        29) Material=HDPE 18   2 yes (0.111111111 0.888888889) *

      15) pipe.age>=11.5 354  33 yes (0.093220339 0.906779661) *
```

This summary of the structure of the decision tree will be explained in detail below:

```
1)          root 2885 532 no (0.815597920 0.184402080)
```

As shown in this summary all observations in training data set that is 2885 observations are represented in the root node. This information provided shows that the majority class of the root node is no, but the 532 observations of the all observations in root are classified as class Yes. Where the root node show that 81.6% (i.e., 0.815597920 as a proportion) of the observations have the target variable pipe break as No, and 18.4% of the observations have it as Yes. The root node is split into two subnodes. The split is based on the variable Pressure_Z.

```
2) Pressure_Z= C1 ,E0.1 ,E0.2, E0.3, Gasil ,Hurush_line, NE1, NE2, NE4, NW0,
NW1+NW2a, NW3+NW2b, out nablus, S2, S3, S4, SE1, SE2, SE3, W-1, W0, W1, W2a,
W2b, W3, W4, Worash_Gravity  2280  174  no (0.923684211 0.076315789)
```

Node 2 has the split expressed as Pressure_Z= C1, E0.1, E0.2, E0.3, Gasil, Hurush_line, NE1, NE2, NE4, NW0, NW1+NW2a ,NW3+NW2b ,out nablus, S2, S3, S4, SE1, SE2, SE3, W-1, W0, W1, W2a, W2b, W3, W4, Worash_Gravity. Node 2 is having 2280 observations of the all observations in root. The majority class of the node 2 is no. Only 174 observations of these 2280 observations are misclassified, which means they are classified as class Yes. the accuracy of predicting for pipe break in the node 2 is 92.4% in predicting that it does not pipe break and 7.6% of the observations have it as Yes.

```
  3) Pressure_Z=Asker_Camps,Ein_Camp,Rafedia,S5,Zawata 605 247 yes (0.408264463
0.591735537)
```

the node 3 has the split expressed as Pressure_Z= Asker_Camps, Ein_Camp, Rafedia, S5, Zawata. The node 3 is having 605 observations of the 2885 observations that are all observations in root. The majority class for this node is yes. the 247 observations of it are represented the pipe won't break, which means the accuracy of predicting for pipe break in the node 3 is 59.2% of pipe (i.e., 0.591735537 as a proportion) in predicting that it does not pipe break. and 40.8% of the observations have it as class no.

Each one of node 2 and node 3 are split into two subnodes. Node 2 is split into node 4 and node 5. The split is based on the variable pipe.age.

```
4) pipe.age< 49.5 2237 132 no (0.940992401 0.059007599)
```

The majority class for the node 4 is no. which has the split expressed as pipe.age< 49.5 and 2237 observations of all observations in the node 2. The 132 observations of it are misclassified, and the accuracy of predicting for pipe break in it is 94% of pipe that it does not pipe break and 6% of the pipe in predicting that it will pipe break.

```
5) pipe.age>=49.5 43   1 yes (0.023255814 0.976744186) *
```

The majority class for the node 5 is yes. Which has the split expressed a pipe.age>= 49.5. 97.7% of the observations have it as yes. And 2.3% of the observations have it as no. node 5 is marked with an asterisk (*), indicating that they are terminal nodes (they are not further split).

```
    8) Material=DI, GS, HDPE, PVC   1965  70   no (0.964376590 0.035623410)

    9) Material=PP, ST 272  62 no (0.772058824 0.227941176)
```

Node 4 is split into node 8 and node 9. The split is based on the variable Material.

Node 8 when Material= DI,GS,HDPE,PVC the majority class is no. 3.6% of the observations have it as yes. and 96.4% of the observations have it as no. node 8 is split into node 16 and node 17. The split is based on the variable Pressure_Z.

Node 9 when Material= PP, ST the majority class is no. 22.8% of the observations have it as yes. and 77.2% of the observations have it as no.
Node 9 is split into node 18 and node 19. The split is based on the variable pipe.age.

```
      16) Pressure_Z= E0.1, E0.3, Gasil,Hurush_line, NE2, NE4, NW0, NW1+NW2a,
NW3+NW2b, out nablus, S2, SE1, SE2, SE3, W2b, Worash_Gravity 925   7    no
(0.992432432 0.007567568) *
```

```
      17) Pressure_Z= C1, E0.2, NE1, S3, S4, W-1, W0, W1, W2a, W3, W4  1040    63
no     (0.939423077 0.060576923)
```

Node 16 when Pressure_Z= E0.1, E0.3, Gasil,Hurush_line, NE2, NE4, NW0, NW1+NW2a, NW3+NW2b, out nablus, S2, SE1, SE2, SE3, W2b, Worash_Gravity the majority class is no. 0.7% of the observations have it as yes, and 99.3% of the observations have it as no. they are terminal nodes

Node 17 when Pressure_Z=C1,E0.2,NE1,S3,S4,W-1,W0,W1,W2a,W3,W4 the majority class is no. 0.6% of the observations have it as yes, and 94% of the observations have it as no.

Node 17 is split into node 34 and node 35. The split is based on the variable type.

```
        34) Type=dis_1 435   6 no (0.986206897 0.013793103) *

        35) Type=dis_2 605  57 no (0.905785124 0.094214876)
```

Node 34 when Type=dis_1 the majority class is no. 1.4% of the observations have it as yes, and 98.6% of the observations have it as no. they are terminal nodes

Node 35 when Type=dis_2 the majority class is no. 9.4% of the observations have it as yes. and 90.6% of the observations have it as no.

Node 35 is split into node 70 and node 71. The split is based on the variable Material.

```
          70) Material= GS, PVC 582  36   no (0.938144330 0.061855670) *

         71) Material= HDPE 23   2    yes (0.086956522 0.913043478) *
```

Node 70 when Material= GS, PVC the majority class is no. 6.2% of the observations have it as yes, and 93.8% of the observations have it as no. they are terminal nodes

Node 71 when Material=HDPE the majority class is yes. 8.7% of the observations have it as no. and 91.3% of the observations have it as yes, they are terminal nodes

```
18) pipe.age< 12.5 205   5 no (0.975609756 0.024390244) *

19) pipe.age>=12.5 67  10 yes (0.149253731 0.850746269) *
```

Node 18 when pipe.age< 12.5 the majority class is no. 2.4% of the observations have it as yes. and 97.6% of the observations have it as no. and they are terminal nodes
Node 19 when pipe.age>=12.5 the majority class is yes. 14.9% of the observations have it as no. and 85.1% of the observations have it as yes. they are terminal nodes

Refer to node 3, They are split into two nodes, and they are node 6 and node 7
The split is based on the variable Diam

```
6) Diam>=95 208  21 no (0.899038462 0.100961538)

7) Diam< 95 397  60 yes (0.151133501 0.848866499)
```

Node 6 when Diam>=95 the majority class is no. 10.1% of the observations have it as yes, and 89.9% of the observations have it as no.
Node 6 is split into node 12 and node 13. The split is based on the variable pipe.age.

```
12) pipe.age< 35 189   3 no (0.984126984 0.015873016) *

13) pipe.age>=35 19   1 yes (0.052631579 0.947368421) *
```

Node 12 when pipe.age< 35 the majority class is no. 1.6% of the observations have it as yes. and 98.4% of the observations have it as no. and they are terminal nodes

Node 13 when pipe.age>=35 the majority class is yes. 5.3% of the observations have it as no. and 94.7% of the observations have it as yes. they are terminal nodes

Node 7 when Diam< 95 the majority class is yes. 94.7% of the observations have it as yes. And 5.3% of the observations have it as no.
Node 7 is split into node 14 and node 15. The split is based on the variable pipe.age.

```
14) pipe.age< 11.5 43  16 no (0.627906977 0.372093023)

15) pipe.age>=11.5 354  33 yes (0.093220339 0.906779661) *
```

Node 14 when pipe.age< 11.5. The majority class is no. 62.8% of the observations have it as no. and 37.2% of the observations have it as yes. Node 14 is split into node 28 and node 29. The split is based on the variable Material.

Node 15 when pipe.age>=11.5. The majority class is yes. 9.3% of the observations have it as no. and 90.7% of the observations have it as yes. They are terminal nodes

```
28) Material=GS, PP, ST 25   0 no (1.000000000 0.000000000) *

29) Material=HDPE 18   2 yes (0.111111111 0.888888889) *
```

Node 28 when Material=GS, PP, ST the majority class is no. 100% of the observations have it as no. and they are terminal node.
Node 29 when Material=HDPE the majority class is yes. 11.1% of the observations have it as no. and 88.9% of the observations have it as yes. And they are terminal node

A.2 Decision Tree as rules:

Rule number: 5 [pipe.break=yes cover=43 (1%) prob=0.98]

Pressure_Z=C1,E0.1,E0.2,E0.3,Gasil,Hurush_line,NE1,NE2,NE4,NW0,NW1+NW2a,NW3+NW2b,out nablus,S2,S3,S4,SE1,SE2,SE3,W-1,W0,W1,W2a,W2b,W3,W4,Worash_Gravity

pipe.age>=49.5

Rule number: 13 [pipe.break=yes cover=19 (1%) prob=0.95]

Pressure_Z=Asker_Camps,Ein_Camp,Rafedia,S5,Zawata

Diam>=95

pipe.age>=35

Rule number: 71 [pipe.break=yes cover=23 (1%) prob=0.91]

Pressure_Z=C1,E0.1,E0.2,E0.3,Gasil,Hurush_line,NE1,NE2,NE4,NW0,NW1+NW2a,NW3+NW2b,out nablus,S2,S3,S4,SE1,SE2,SE3,W-1,W0,W1,W2a,W2b,W3,W4,Worash_Gravity

pipe.age< 49.5

Material=DI,GS,HDPE,PVC

Pressure_Z=C1,E0.2,NE1,S3,S4,W-1,W0,W1,W2a,W3,W4

Type=dis_2

Material=HDPE

Rule number: 15 [pipe.break=yes cover=354 (12%) prob=0.91]

Pressure_Z=Asker_Camps,Ein_Camp,Rafedia,S5,Zawata

Diam< 95

pipe.age>=11.5

Rule number: 29 [pipe.break=yes cover=18 (1%) prob=0.89]

Pressure_Z=Asker_Camps,Ein_Camp,Rafedia,S5,Zawata

Diam< 95

pipe.age< 11.5

Material=HDPE

Rule number: 19 [pipe.break=yes cover=67 (2%) prob=0.85]

Pressure_Z=C1,E0.1,E0.2,E0.3,Gasil,Hurush_line,NE1,NE2,NE4,NW0,NW1+NW2a,NW3+NW2b,out nablus,S2,S3,S4,SE1,SE2,SE3,W-1,W0,W1,W2a,W2b,W3,W4,Worash_Gravity

pipe.age< 49.5

Material=PP,ST

pipe.age>=12.5

Rule number: 70 [pipe.break=no cover=582 (20%) prob=0.06]

Pressure_Z=C1,E0.1,E0.2,E0.3,Gasil,Hurush_line,NE1,NE2,NE4,NW0,NW1+NW2a,NW3+NW2b,out nablus,S2,S3,S4,SE1,SE2,SE3,W-1,W0,W1,W2a,W2b,W3,W4,Worash_Gravity

pipe.age< 49.5

Material=DI,GS,HDPE,PVC

Pressure_Z=C1,E0.2,NE1,S3,S4,W-1,W0,W1,W2a,W3,W4

Type=dis_2

Material=GS,PVC

Rule number: 18 [pipe.break=no cover=205 (7%) prob=0.02]

Pressure_Z=C1,E0.1,E0.2,E0.3,Gasil,Hurush_line,NE1,NE2,NE4,NW0,NW1+NW2a,NW3+NW2b,out nablus,S2,S3,S4,SE1,SE2,SE3,W-1,W0,W1,W2a,W2b,W3,W4,Worash_Gravity

pipe.age< 49.5

Material=PP,ST

pipe.age< 12.5

Rule number: 12 [pipe.break=no cover=189 (7%) prob=0.02]

Pressure_Z=Asker_Camps,Ein_Camp,Rafedia,S5,Zawata

Diam>=95

pipe.age< 35

Rule number: 34 [pipe.break=no cover=435 (15%) prob=0.01]

Pressure_Z=C1,E0.1,E0.2,E0.3,Gasil,Hurush_line,NE1,NE2,NE4,NW0,NW1+NW2a,NW3+NW2b,out nablus,S2,S3,S4,SE1,SE2,SE3,W-1,W0,W1,W2a,W2b,W3,W4,Worash_Gravity

pipe.age< 49.5

Material=DI,GS,HDPE,PVC

Pressure_Z=C1,E0.2,NE1,S3,S4,W-1,W0,W1,W2a,W3,W4

Type=dis_1

Rule number: 16 [pipe.break=no cover=925 (32%) prob=0.01]

Pressure_Z=C1,E0.1,E0.2,E0.3,Gasil,Hurush_line,NE1,NE2,NE4,NW0,NW1+NW2a,NW3+NW2b,out nablus,S2,S3,S4,SE1,SE2,SE3,W-1,W0,W1,W2a,W2b,W3,W4,Worash_Gravity

pipe.age< 49.5

Material=DI,GS,HDPE,PVC

Pressure_Z=E0.1,E0.3,Gasil,Hurush_line,NE2,NE4,NW0,NW1+NW2a,NW3+NW2b,out nablus,S2,SE1,SE2,SE3,W2b,Worash_Gravity

Rule number: 28 [pipe.break=no cover=25 (1%) prob=0.00]

Pressure_Z=Asker_Camps,Ein_Camp,Rafedia,S5,Zawata

Diam< 95

pipe.age< 11.5

Material=GS,PP,ST

TableTableTableTable

ملخص البحث

تعاني العديد من الدول في العالم من نقص في المياه, مما يتوجب ضرورة توفير المياه و ادارتها بشكل صحيح بالاضافة الى ايجاد المشاكل التي تسبب فقدان المياه و معالجتها.

و تعد مشكلة فقدان المياه من شبكات توزيع المياه من مسببات نقص المياه, حيث أنها تشكل مشكلة اجتماعية واقتصادية كبيرة تثير قلق العديد من الدول في العالم بشكل عام, وخاصة العاملين في البلديات المسؤولة عن شبكات توزيع المياه.

حيث أن كسر الأنابيب في شبكات توزيع المياه هو أحد الأسباب الرئيسية لفقدان المياه من الشبكة، لذلك هناك حاجة ملحة للسيطرة على هذه المشكلة ومحاولة منع تسرب المياه من الأنابيب عن طريق محاولة إصلاح وصيانة الأنابيب قبل أن يكسر. لذلك، هناك حاجة لتحليل وفهم البيانات المتعلقة بشبكات توزيع المياه واستخدام هذه البيانات في التنبؤ بكسر خطوط الأنابيب وتحديد العوامل والمتغيرات التي تؤدي إلى كسرها.

ولأن استخدام الأدوات الرياضية والإحصائية الكلاسيكية في تحديد المتغيرات التي تلعب دورا رئيسيا في التنبؤ بأنماط انقطاع الأنابيب عملية معقدة؛ يسعى هذا البحث إلى إيجاد نموذج بديل يستخدم لتنبؤ بكسر الانابيب الرئيسية في شبكات توزيع المياه، وتحديد المتغيرات التي تسبب هذه الفواصل, باستخدام تقنيات تنقيب البيانات و لتحقيق ذلك تم استخدام نظام توزيع المياه في بلدية نابلس كدراسة حالة, و التي تعد من المدن الكبرى في منطقة شمال الضفة الغربية في فلسطين.

استخدمت لغة R في هذا البحث لتنفيذ تقنيات تنقيب البيانات, حيث تم استخدام ثلاثة تقنيات معروفة هي شجرة القرارات و الانحدار اللوجستي و شعاع الدعم الالي لبناء نماذج تصنيف ثنائية , حيث تصنف الانابيب الى صنفين حسب حالة الانبوب (يتوقع كسره او لا) . بالاضافة الى بناء نماذج من خلال الدمج بين هذه التقنيات مما نتج عنها سبع نماذج مختلفة. و كنتيجة لتقييم أداء جميع النماذج التي تم بناؤها تبين ان النموذج الذي تم بناؤه من خلال الدمج بين نموذجي الانحدار اللوجستي و شعاع الدعم الالي هو أفضل نموذج للتنبؤ بكسر الانابيب؛ حيث أعطى أفضل القيم لمعظم مقاييس الأداء المحسوبة حيث تراوح معدل الخطأ بين 0.01 و 0.12، كما أنه قادر على توفير ما يصل إلى 0.97 المياه من كمية المياه المفقودة من الشبكة، مع معدل دقة قد يصل إلى 0.99.