Journal of Quality Engineering and Production Optimization

Vol. 8, No. 1, Winter & Spring 2023

http://jqepo.shahed.ac.ir

Research Paper

DOI: 10.22070/jqepo.2023.16409.1237

Predicting Project Delays Using New Trended Regression Tree Method

Mohamad Ali Movafaghpour[1*]

[1] *Jundi-Shapur University of Technology, Dezful, Iran*

*** Corresponding Author:*** *Mohamad Ali Movafaghpour (Email: Movafaghpour@jsu.ac.ir)*

Abstract *–Accurate prediction of potential delays in pipeline projects can provide valuable information relevant for mitigating completion risk in future natural gas distribution projects. However, existing techniques for evaluating completion risk remain incapable of identifying hidden patterns in risk behavior within the vast database of projects. The purpose of this paper is to model project delays. Sample instances are drawn from the database of recent natural gas distribution projects in Iran between 2015 and 2020. A series of predictive models have been reviewed and evaluated for delay risk prediction, such as k-Nearest Neighbor (k-NN) Regression, Regression Trees (RT), Support Vector Machine Regression (SVMR), and Artificial Neural Network (ANN). Computational results based on cross-validation revealed that when delays follow a rational pattern, they can be predicted by our developed Trended Regression Tree (TRT) method and k-NN regression method. These novel methods are effective and provide practitioners with significantly more reliable predictions and applied insight into the delay causes. The concept of Trended Regression Trees is developed for the first time. Project delays are modeled based on project specifications, and therefore there is no need to make any extra data gathering to predict project delays. Based on the research findings, we recommended that the management team focus on the most effective factors to reduce project delays.*

Keywords*– Prediction Model, Project Delay Factors, Classification and Regression Tree (CART), Natural Gas Distribution Projects.*

I. INTRODUCTION

Oil and gas are the most important energy carriers in the global energy portfolio, which, according to official reports, will continue to play a significant role in the coming decades. Meanwhile, the process of converting natural gas into thermal energy produces less environmental pollution compared to other fossil energy carriers. To continue the process of industrial and economic development of the world while protecting the environment, the development and expansion of natural gas as a clean energy source compatible with environmental considerations are inevitable requirements and prerequisites for sustainable development. In addition to the environmental benefits of using natural gas, factors such as better geographical dispersion and competitive pricing with other energy carriers have led to the success and allocation of an increasing share of natural gas in the global energy basket. In Iran, with the implementation of over 350,000 kilometers of distribution network, the share of natural gas in the country's energy basket has

Manuscript Received:7-October-2022 & Revised: 11-January-2023 & Accepted: 1-March-2023
ISSN: 2423-3781

reached more than 70%. The implementation of natural gas distribution projects has always been associated with delays, so the study in identifying and modeling delays can help the country's economy save costs and improve the national welfare.

Time delays are important in a wide variety of engineering fields and even human behavioral sciences. Because such delays follow complicated patterns, it is difficult to model them with conventional tools. E.g., Cui et al. (2022) developed an Artificial Neural Network to predict ignition delays in combustion engines; Borovsky et al. (2021) employed machine learning techniques to predict language delay as a common language disorder in preschool and school-age children. Predicting project delays with novel machine learning techniques is a promising trend; for example, Sanni-Anibire et al. (2022) used K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), Support Vector Machines (SVM) and Ensemble methods for assessing delay risks in tall buildings. Gurgun et al. (2022) and Egwim et al. (2021) applied artificial intelligence for predicting delays in construction projects. Some other researchers have attempted to forecast project cost or project liquidity using machine learning techniques, such as Shoar et al. (2022), Alshboul et al. (2022a), and Alshboul et al. (2022b). Furthermore, other studies have sought to to predict time and cost risks in construction projects simultaneously, such as Banerjee Chattapadhyay et al. (2021) and Mahmoodzadeh et al. (2022).

Kassem et al. (2021) investigated the causes of risks in oil and gas construction projects in Yemen. They found 51 critical factors that cause risks and divided them into two major groups: internal (including seven sources) and external (including six sources) risk factors. Derakhshanfar et al. (2020) surveyed 118 delayed construction projects in Australia to determine the most impactful delay risks. They identified the five most impactful delay risks as “changes by the owner”, “slow decisions by the owner”, “preparation and approval of design drawings”, “underestimation of project complexity” and “unrealistic duration imposed on the project”. Türkakin et al. (2020) emphasized that schedules for many smaller projects in developing countries, are not updated during construction. In such cases that updated schedules are unavailable, it is not possible to determine which activities caused project delay. Therefore, they developed a method to estimate the share of each activity in the project delay by using the as-planned schedule and the expense logs kept on site. This research faces a similar situation, in which schedule updates during construction are not available, and the project planner seeks to predict the total project delay based on project specifications.

II. PROBLEM DESCRIPTION

In project-oriented organizations, predicting the real project conditions that occur in the field is a key success factor. The more effective variables we can collect based on records, the more accurate our model of future situations will be.

The main goal of this research is to model the project delays using quantitative indicators of records. This method can be integrated into the company-wide project management software. Traditionally, commercial project management software conducts risk analyses based on activity profiles; however, general analysis using records of grand features of the projects can yield more accurate results.

Natural gas distribution projects usually include the construction of pressure reduction stations for header lines and distribution networks to consumers. Construction of the station, depending on the relevant type and standard, includes the construction of the building, and infrastructure, and mechanical installation of the station. Header lines are usually made of steel and require the construction of Cathodic Protection Stations. The contractors employed in these projects have a formal executive rank. Also, the number of branches is one of the quantitative indicators of distribution projects. Such projects, which are implemented over a wide geographical area, often require official permits from other public service provider companies. The number of these permits can be an effective parameter in making changes in the project. Table I reports the important indicators recorded as the main effective factors influencing project delays in the Khuzestan Gas Company. They were identified through another statistical survey (Mehrabi & Movafaghpour, 2021).

Table I. Factors affecting Project Delays.

Category	Indicator	Notation
Project Specification	Number of Cathodic Protection Station (CPS)	x_8
	Number of branches	x_9
	Project cost	x_7
Owner Factors	Number of permits required	x_6
	Number of legal cases	x_4
	Debt of owner (Billion R)	x_3
	Number of design change	x_2
	Delay of inspection	y_3
Contractor Factors	Rank of contractor	x_1
	Delay of performance	y_1
	Delay of materials	y_2
Force Majeure	Number of unfavorable weather days	x_5
Contractor Factors	Rank of contractor	x_1
	Delay of performance	y_1
	Delay of materials	y_2
Owner Factors	Number of design change	x_2
	Debt of owner (Billion R)	x_3
	Number of legal cases	x_4
	Number of permits required	x_6
	Delay of inspection	y_3
Force Majeure	Number of unfavorable weather days	x_5
	Project cost	x_7
Project Specification	Number of CPS	x_8
	Number of branches	x_9

Permitting and land acquisition are two key factors for progressing pipeline projects from design to construction. Moreover, the number of legal cases faced when acquiring required land for the project is another important factor; therefore, the number of legal cases faced by the project is considered a quantitative index. Execution of natural gas distribution projects is accompanied by continuous inspection and control of consumed goods and how to execute the operation process. Delays in this inspection operation can lead to delays throughout the project. We call this type of delay "inspection delays." Projects that do not run on a limited site are usually affected by weather conditions, making the number of unfavorable days in projects is a key factor in delays. Design changes also occur in pipeline projects due to the lack of anticipation of some underground barriers. Delays related to the contractor inefficiency fall outside the scope of organizational operations and the management of the owner. These factors include poor "site management and supervision by the contractor" (labeled as "Delay of Performance" in Table I) and delayed "Supply of Materials" (labeled as "Delay of Material" in Table I). Related indicators are categorized in Table I.

III. LITERATURE REVIEW

Many researchers have yet to focus on identifying the causes of delays in projects. Understanding these factors helps managers to concentrate on the most important factors and thus reduce the risks of project delays. Project information is often divided into two categories: quantitative and qualitative indicators. These indicators can be directly related to the causes of delays. For example, the official index of the contractor's rank reflects their executive records and financial strength, and therefore can directly predict the risks of delays due to the weakness of the contractor.

Yang and Wei (2010) emphasized the causes of delays during the planning and design stages. Their findings show that the most important reason for the delay in the planning and design phase is a "change in customer needs". Al-Kharashi and Skitmore (2009) reviewed ten previous studies on public projects in Saudi Arabia as a basis for their research. Based on these studies, 112 causes related to several parties and project parameters were identified, including "customer, consultant, contractor, materials, contract, relationships and work". The final results highlighted the most important cause for each group. Chan and Kumaraswamy (2002) tried to assess the importance of delays in two types of construction projects in Hong Kong. Previous research on construction and civil engineering projects identified 83 causes of delays. Based on these results, the authors set up a questionnaire to gain expert opinions on the importance of each cause. Identifying the causes of delays can be studied from different aspects as delays are one of the main sources of conflict in projects. They impose heavy overhead costs on contractors and prevent owners from gaining project benefits, so each party tries to blame the other.Abd El-Razek et al. (2008) investigated the causes of delays and indicated that the most important causes were financed by the contractor during construction, delays in the contractor's payment by the owner, design changes made by the owner or his agent during construction, partial payments during construction, and non-utilization of professional construction/contractual management. Doraisamy et al. (2015) presented an overview of project delays, identifying various causes through their research by different researches and proposing appropriate recommendations to overcome and eliminate problems that hindered the success of the construction projects. Gunduz et al. (2015) surveyed to identify delay factors on construction projects in Turkey; they identified 83 different delay factors and categorized them into nine major groups. Fallahnejad (2013) identified 43 causes of delay and ranked these causes in Iranian gas transmission pipeline projects. The study showed that the ten major delay factors included: imported materials, unrealistic project duration, client-related materials, land expropriation, change orders, contractor selection methods, payment to the contractor, obtaining permits, suppliers, and contractor cash flow. Sambasivan et al. (2017) analyzed delays in Tanzanian construction projects using transaction cost economics; they developed a structural equation model based on 32 identified delay causes grouped into seven categories. Adam et al. (2017) conducted a literature analysis to provide an aggregated ranking of project delays, which was implied to 40 journal articles reporting on delays in publicly-funded construction projects. Islam and Trigunarsyah (2017) in a review paper tried to present the causes and effects of construction delays in developing countries. They collected relevant literature from 28 developing countries through scholarly journals published between 2006 and 2016. Mohammed and Suliman (2019) identified delay factors including poor managerial skill, slow decision-making within all project teams, lack of communication between client, consultant, and contractor; inadequate design team, scope variations, unrealistic contract decisions, and delays in project drawing preparations. They identified the factors causing delays and the associated risks in pipeline construction projects in Bahrain. Durdyev and Hosseini (2020) presented a systematic review of construction project delays published between 1985 and 2018, highlighting that most research originates from developing countries.

Project evaluation and selection using fuzzy group judgments (Davoudabadi et al., 2019 ; Gitinavard & Mousavi, 2015) and incomplete information (Gitinavard, 2019) have been the focus of several studies. Gitinavard et al. (2020)

developed a weighting method for safety evaluation under a hesitant fuzzy environment. To determine the validity of their proposed model, they compared the performance of their proposed method with a method developed by Zhang and Wei (2013), which extended the TOPSIS and VIKOR methods in a hesitant fuzzy environment. Finally, both methods resulted in the same ranking results.

Lin and Fan (2019) explored the capabilities of three kinds of decision tree algorithms, namely Classification and Regression Tree (CART), Chi-Squared Automatic Interaction Detection (CHAID), and Quick Unbiased Efficient Statistical Tree algorithms (QUEST), in predicting construction project defects.

Ilic et al. (2021) proposed an Explainable Boosted Linear Regression (EBLR) algorithm for time series forecasting, which is an iterative method that starts with a base model and explains the model's errors through regression trees.

This study seeks to predict project completion delays based on records. Traditional techniques for function estimation based on given values of a set of independent variables consist of Artificial Neural Network (ANN), Genetic Programming (GP), and Regression.

Regression analysis is selected from several samples following the estimation of the relationships between output variables and a set of independent input variables with automatic learning (Sen & Srivastava, 2012). The main purpose of regression analysis is usually to achieve an appropriate prediction of the level of output variables for new samples. Examples of regression analysis methods in the literature include linear regression (Seber & Lee, 2012), automatic learning of algebraic models for optimization (ALAMO) (Cozad et al., 2014; Zhang & Sahinidis, 2013), Support Vector Regression (SVR) (Smola & Scholkopf, 2004), k-Nearest Neighbor (k-NN) (Korhonen & Kangas, 1997), Multivariate Adaptive Regression Lines (MARS) (Friedman, 1991; Kleijnen, 2017), and the regression tree. Often, we want to gain a useful understanding of the relationship between input and output variables, making interpretability of the regression method significant.

A regression tree is a type of machine learning tool that can meet good predictive accuracy and easy interpretation, thus being widely considered in the literature. The regression tree uses a tree diagram or model, which is created in an iterative process that divides each node into child nodes by specific rules, unless it is an end node where samples are placed. To obtain the predicted values of the output variables of the new samples, a regression model is set for each terminal node.

The Classification and Regression Tree (CART) is probably the most famous decision tree learning algorithm in the literature (Breiman et al.,1983). Depending on the set of samples, the CART specifies an input variable and a breakpoint before dividing the samples into two child nodes. Starting from the existing instruction set (root node), a return binary partitioning occurs for each node until further division is possible or specific termination criteria are met. In each node, the best division is determined by a comprehensive search, i.e. all possible gaps on each input variable and each breakpoint are tested. This process identifies the division that results in the minimum deviations by predicting the outcomes for the two child nodes of the sample. They were selected by the average of their output variables. After the tree growth method, an overgrown tree is usually created, which leads to a lack of generalization of the model to unseen specimens. The pruning method is used to remove gaps that contribute improperly to the accuracy of the exercise. This tree is cut from the maximum size tree to the end of the root node, resulting in a sequence of candidate trees. Each candidate tree is then tested with an unseen validation dataset, with the tree exhibiting the lowest prediction error selected as the final tree (Breiman, 2001). Conventionally all regression tree algorithms assign a constant prediction to each terminal node (Loh, 2011) to minimize misclassification costs. A classification tree algorithm called CRUISE can optionally fit bivariate linear discriminant models in the nodes (Loh, 2011).

Recently Hamzeh et al. (2020) focused on project duration risk and proposed a Triangular Intuitionistic Fuzzy Earned Duration Management model to forecast time performance of projects under uncertain conditions. They introduced the notions of non-membership and hesitation degrees to define time-based risk performance indicators. One of the best recent tools to tackle complex and non-structured regression problems is Cubist (RuleQuest, 2016), a commercially available rule-based regression model that has gained increasing popularity recently (Yang et al., 2017). This research introduces a regression tree algorithm that hierarchically splits observations into a tree-like structure and then fits a robust multiple regression model on each terminal node.

Li et al. (2022) focused on predicting ambulance delays when transferring a patient to a hospital. They used a Naive Bayes classifier to remove the noisy training observations and then developed a basic Classification and Regression Tree (CART) algorithm. Mittas and Mitropoulos (2022) proposed using CART and KNN for predicting construction cost of natural gas pipeline projects. Ghazal and Hammad (2022) identified twelve factors affecting cost overrun of construction projects. They developed several data mining tools and their method predicted cost overrun of construction projects with less than 61 percent of accuracy. Alshboul et al. (2022) used *k-NN*, Decision Tree and several other machine learning techniques for predicting delays in highway construction projects. Finally, they concluded that Machine Learning techniques could be used as effective administrative decision adding tools for forecasting performance measures of building projects. Taleongpong et al. (2021) developed a gradient boosting algorithm and several other machine learning techniques to predict delays occurred in British railway network caused by chain reaction delays. The same problem and similar set of algorithms was used by Klumpenhouwer and Shalaby (2022) to investigate a rail network in Canada.

IV. SOLUTION APPROACH

A. Trended Regression Trees

Regression trees are helpful tools for decision support and predictive analytics due to their simple structure and the ease with which they can be obtained from data. The resulting Regression Tree (RT) looks like a hierarchical clustering scheme. Each node is split into two branches based on a threshold for the value of a distinct variable (dimension) of the observations. Fig.1 depicts a sample binary tree. Each parent node may be divided into two parent nodes or two terminal nodes (leaves).

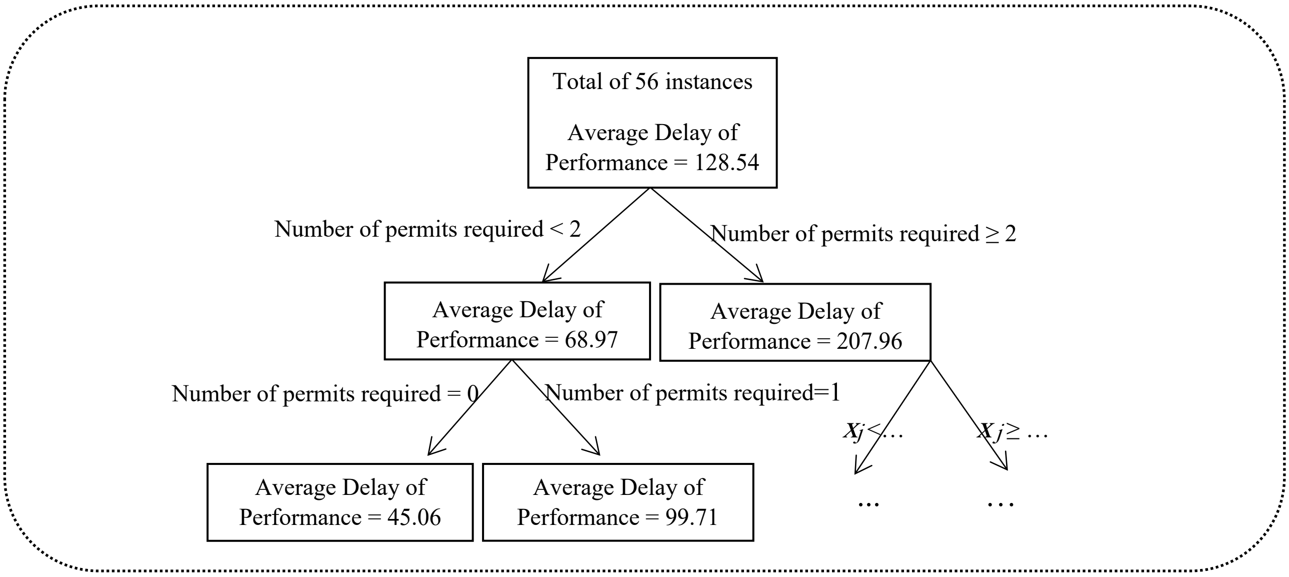
Total of 56 instances
Average Delay of Performance = 128.54
Number of permits required < 2
Number of permits required ≥ 2
Average Delay of Performance = 68.97
Average Delay of Performance = 207.96
Number of permits required = 0
Number of permits required=1
X_j< …
X_j≥ …
Average Delay of Performance = 45.06
Average Delay of Performance = 99.71
…
…

Fig. 1. A schematic view of a sample binary splitting on the observations of the project delay database

Creating an RT model involves selecting input variables and splitting data points based on those variables until a suitable tree is constructed. The selection of which input variable to use and the specific split or cut-point is chosen using a greedy algorithm to minimize an error function, such as Mean Squared Error (MSE), Root Mean Squared Error

(RMSE), or Mean Absolute Error (MAE). Tree construction ends using a predefined stopping criterion, such as tree depth or a minimum number of training instances assigned to each parent or leaf node of the tree.

Creating a binary RT is a recursive process of dividing up the input instances. In this numerical procedure, at each node, all instances are sorted with regard to each variable (dimension), and the split point is slipped increasingly through the values of the variables of interest. The best split point is identified where the error function meets its minimum value. All input variables and all possible split points are evaluated and the best one is chosen in a greedy manner.

Traditional RT concludes with a binary tree that at each of its terminal leaf *c*, a value of $\overline{y}_c$ represents the average output variable for all the instances assigned to that leaf. Only averaging the output variables $y_i, (i \in c)$ may ignore some useful information behind the values of input variables. $x_{i,m}, (i \in c)$. Therefore, we prefer to develop a linear regression function at each terminal leaf. Specifically, after the full-size RT is produced, we perform a linear independency analysis among *x* variables at each leaf to distinguish the independent variable set based on the instances assigned to each leaf. This analysis is done because some leaves may have *C* instances, and we require *C* >*M* to fit a linear regression function of the form:

$$\hat{y}_i = a_m x_{i,m} + a_{i,m-1} x_{i,m-1} + \ldots + a_0 \tag{1}$$

The overall pseudo code of the Trended Regression Tree construction algorithm used in this research is described as follows:

Step 1. Start with a single node that includes all data points. Each *i*-th data point entails a set of *J*-Dimensional independent variables x_{ij} (*j* = 1, … *J*), and a dependent variable y_i.

Step 2. For each *j*-th dimension in the current node *c*:

Sort all data points *i* (i = 1, …, *I*) in the current node *c* according to the value of their *j*-th dimension.

Step 3. For each *k*-th data point in the current node *c*:

Calculate the progressive error S_{kj} to find the best split point S_{k^*j} ($1 \le k \le I$):

$$S_{kj} = \sum_{x_{ij} \le x_{kj}} \left(y_i - \overline{y_k}\right)^2 + \sum_{x_{ij} > x_{kj}} \left(y_i - \overline{y_{k'}}\right)^2 \tag{2}$$

where

$$\overline{y_k} = \frac{1}{k} \sum_{x_{ij} \le x_{kj}} y_i \, , \; \overline{y_{k'}} = \frac{1}{I-k} \sum_{x_{ij} > x_{kj}} y_i \tag{3}$$

Step 4. Find the lowest value S_{k^*j} and partition the data points of the current node into two groups: those data points before *k**, and those after it. Name each partition as a child node.

Step 5. If there is a node with data points more than the threshold, set that as the current node and go to Step 2. Otherwise, stop.

In order to fit a linear regression model at each node, it is required to ignore dimensions with collinearity. Making a linear independency analysis among the dimensions is done by performing a Gauss-Jordan elimination to find the Reduced Row Echelon Form. By ignoring dependent columns, a linear regression model is fitted with regard to independent columns. Fig.3 depicts a sample Trended Regression Tree (TRT).

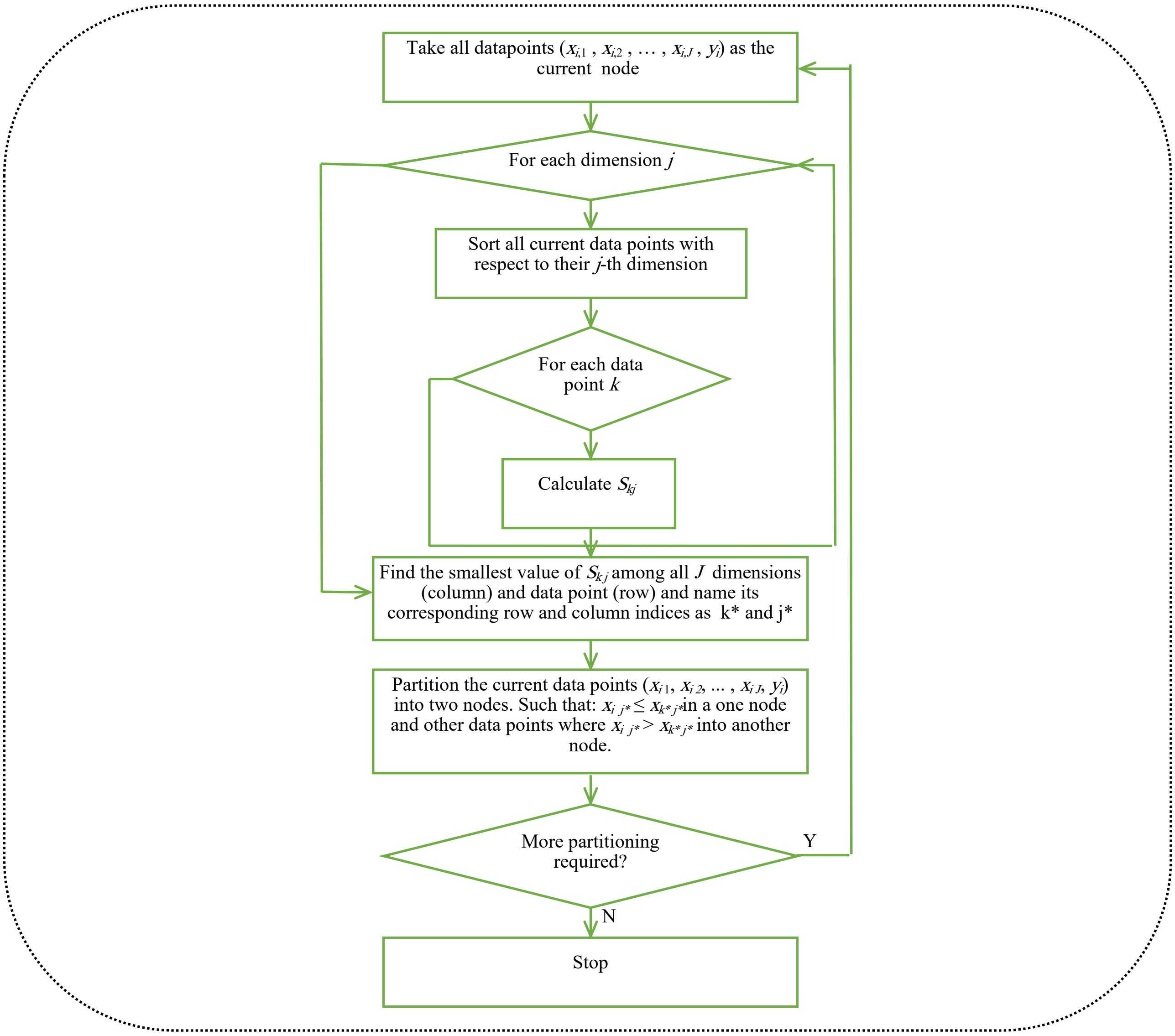
Take all datapoints ($x_{i,1}$, $x_{i,2}$, … , $x_{i,J}$, y_i) as the current node
For each dimension j
Sort all current data points with respect to their j-th dimension
For each data point k
Calculate S_{kj}
Find the smallest value of S_{kj} among all J dimensions (column) and data point (row) and name its corresponding row and column indices as k* and j*
Partition the current data points (x_{i1}, x_{i2}, … , x_{iJ}, y_i) into two nodes. Such that: $x_{ij*} \leq x_{k*j*}$ in a one node and other data points where $x_{ij*} > x_{k*j*}$ into another node.
More partitioning required?
Y
N
Stop

Fig. 2. A schematic view of binary tree generation procedure.

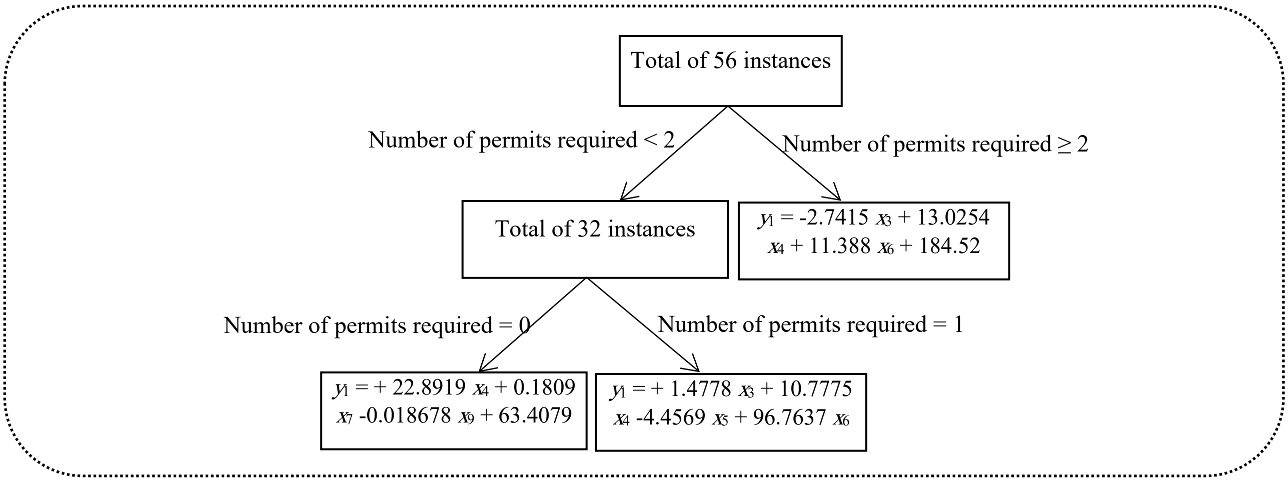
Total of 56 instances
Number of permits required < 2
Number of permits required ≥ 2
Total of 32 instances
y_1 = -2.7415 x_3 + 13.0254 x_4 + 11.388 x_6 + 184.52
Number of permits required = 0
Number of permits required = 1
y_1 = + 22.8919 x_4 + 0.1809 x_7 -0.018678 x_9 + 63.4079
y_1 = + 1.4778 x_3 + 10.7775 x_4 -4.4569 x_5 + 96.7637 x_6

Fig.3. A compact view of a sample Trended RT on the observations of project delay database.

B. k-NN Regression

A simple implementation of *k*-NN regression is to calculate the average of the numerical output variable of the *k* nearest neighbors as an estimate of the unseen output variable:

$$\hat{y}_j = \frac{1}{k} \sum_{l \in N(j)} y_l \tag{4}$$

Nearest neighbors are recognized based on several metrics in the literature such as Euclidian, Manhattan, Minkowski, or Hamming distance. Without any lack of generality, this paper uses Euclidian distance measure here as:

$$distance\left(project_i, project_j\right) = \sqrt{\sum_m \left(x_{i,m} - x_{j,m}\right)^2} \tag{5}$$

One major drawback in calculating distance measures directly from the raw input data arises when variables have different measurement scales or when there is a mixture of numerical and categorical variables. For example, in our database of gas distribution pipelines, the number of Cathodic Protection Stations (x_1) ranges from 0 to a maximum of 3, while the number of branches in each gas distribution project (x_2) ranges from 400 to 3700; different ranges for variables will have a bias effect on distance metric because the much higher influence on the distance is affected by variables with higher values. One solution is to standardize the values of each variable into the range of 0 to 1 as below:

$$x'_{i,j} = \frac{x_{i,j} - \min_j \left\{x_{i,j}\right\}}{\max_j \left\{x_{i,j}\right\} - \min_j \left\{x_{i,j}\right\}} \tag{6}$$

Although using a simple arithmetic mean with equal weight results in less computational effort and is the dominant procedure in the literature but this kind of simplistic estimation ignores the hidden patterns behind the potential trend of variables. Therefore, we prefer to develop a linear regression function $f\left(x_{i,1}, x_{i,2}, \ldots, x_{i,m}\right)$ to predict the dependent variable y_i as below:

$$\hat{y}_i = a_m x_{i,m} + a_{i,m-1} x_{i,m-1} + \ldots + a_0 \tag{7}$$

To fit a linear regression function and predict the project delay, we can use each one of two potential approaches:

1) Constructing neighborhood clusters based on training data instances and fitting a linear regression function over each of these “Early Clusters” (EC) learned from training data.
2) Recognizing the *k* nearest neighbors for each unseen (test) project instance among all given training project instances, then fitting a linear regression function over this “Recent Cluster” (RC).

In the remainder of this manuscript, we will refer these two kinds of regression functions as “Early Cluster” (EC) and “Recent Cluster” (RC), respectively.

In both EC and RC, if the predicted output value generated by the regression function lies outside the interval bounded by the minimum and maximum of the output values for training samples in that neighborhood cluster, is then adjusted to the nearest bound.

It is obvious that each unseen project instance can be mapped into a single cluster formed by its *k* nearest neighbors. Therefore, we have a single estimate $\hat{y}_i$ as RC for each unseen instance *i*.

In contrast to the single estimate RC regression line for each unseen instance, each of the *k* nearest neighbors of each unseen project instance has *k* regression lines related to *k* EC's. Therefore, we have *k* estimates $\hat{y}_i(r), r = 1, 2, \ldots, k$ as EC for each unseen instance *i*. To integrate the *k* given EC's into a single estimate $\hat{y}_i$, the following consensus-making equations are proposed:

$$\hat{y}_i = \underset{r}{median}\left(\hat{y}_i(r)\right) \tag{8}$$

$$\hat{y}_i = \frac{1}{2}\left(percentile\left(\hat{y}_i(r), 25\right) + percentile\left(\hat{y}_i(r), 75\right)\right) \tag{9}$$

$$\hat{y}_i = \frac{1}{2}\left(percentile\left(\hat{y}_i(r), 30\right) + percentile\left(\hat{y}_i(r), 70\right)\right) \tag{10}$$

Fig. 3 depicts the idea of EC's formed around each data point in the training dataset, along with a single RC formed by gathering *k* nearest neighbors of an unseen data point.

When a set of *k* nearest neighbors of a test data point is identified, if the nearest one has a zero distance from the test data point, the other distant neighbors are ignored. This definition of 'Zero Distance Exclusive Neighborhood' causes zero error for seen data and has a good effect on unseen data.

In order to avoid overfitting of the regression line in both methods (*k*–NN Regression and TRT), a correlation analysis is done and less correlated.

Variables with the output variable are ignored, i.e. the input variables that have an absolute value of the correlation coefficient with response variable below the 65th percentile of all correlation coefficients are ignored.

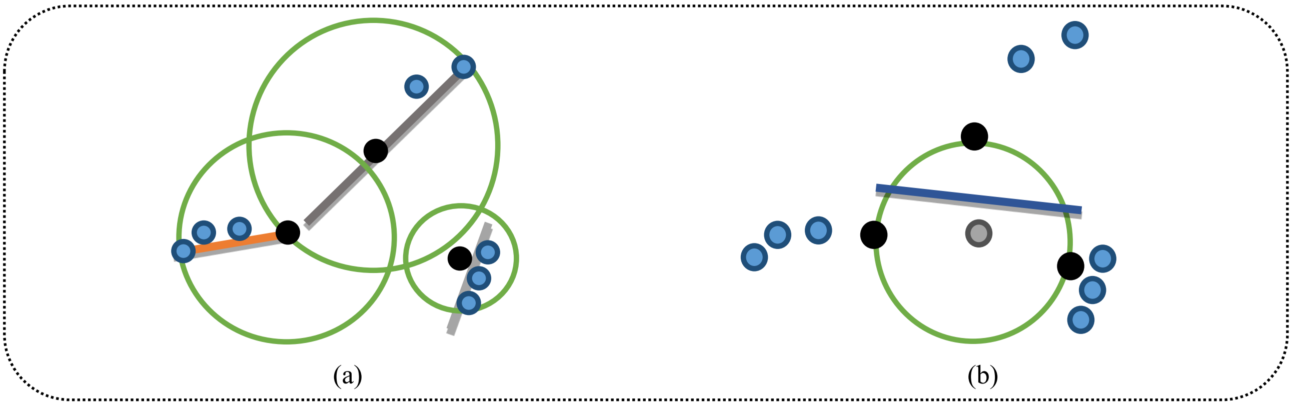

Fig.4. Linear regression lines for *k* = 3: (a) Three "Early Clusters" formed regarding to three black data points in the training dataset. (b) One "Recent Cluster" formed for a new unseen data point in green over the same training data points.

V. COMPUTATIONAL EXPERIMENTS

To predict delays in gas distribution pipeline projects, a database of 56 projects was selected as the most recent projects (from 2015 to 2020) among the company-wide archive of projects. The main cause of focusing on recent projects is to assure the homogeneity of data. As shown in Table I, each project involves 9 input variables as project characteristics denoted as x_i, (where i = 1, …, 9), and 3 kinds of project delay, represented as output variables y_i, (where i = 1, 2, 3). TRT and k-NN regression algorithms are coded in MATLAB. All the computational experiments are implemented on a CORE i7 laptop running Windows 7.

A. Performance measures

In order to evaluate the performance of our developed TRT and k-NN regression algorithms, Artificial Neural Network (ANN), Support Vector Machine (SVM), and CUBIST algorithms are used as standard regression algorithms. The ANN and SVM toolboxes are used from MATLAB, while the CUBIST package (Rule Quest, 2016) runs in R (R Development Core Team, 2008) environment.

B. Efficiency Comparison

The TRT algorithm is implemented in MATLAB. After constructing the binary splitting tree, a correlation analysis is done and the dimensions (variables) are less correlated to the output variable and dimensions (variables) with collinearity are ignored. Finally, a linear regression function is fitted. We set Minimum Parent Size for constructing the RT at 30.

In the remaining of this paper, the k-NN algorithm based on recent clusters is denoted as k–NN RC. The k–NN algorithm based on the median (Eq.6) and other percentile derivatives (Eq.7-8) are denoted as k–NN EC – Median, k–NN EC – 45-55, and k–NN EC – 40-60, respectively. The k factor for all four k-NN regression algorithms is set to 10. We observed that by increasing k from 2 to 10, the accuracy of all four k-NN regression algorithms is improved; however, further increases neighborhood size from 10 up to higher values do not improve the accuracy of k-NN regression algorithms significantly.

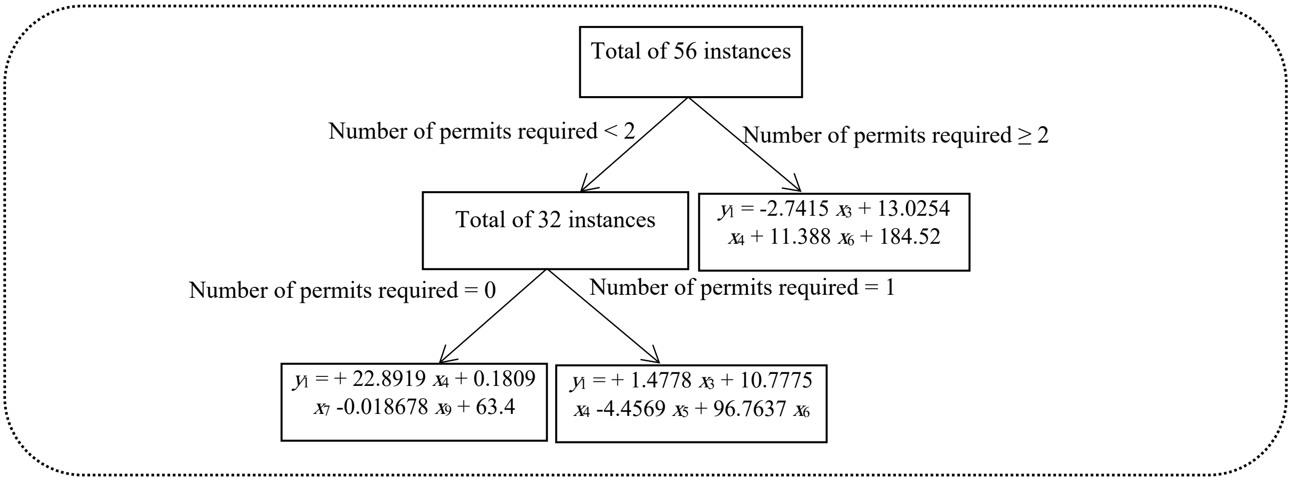
Total of 56 instances
Number of permits required < 2
Number of permits required ≥ 2
Total of 32 instances
y_1 = -2.7415 x_3 + 13.0254 x_4 + 11.388 x_6 + 184.52
Number of permits required = 0
Number of permits required = 1
y_1 = + 22.8919 x_4 + 0.1809 x_7 -0.018678 x_9 + 63.4
y_1 = + 1.4778 x_3 + 10.7775 x_4 -4.4569 x_5 + 96.7637 x_6

Fig. 5. Final view of a Trended RT for predicting Delay of Performance (y_1).

We made an Artificial Neural Network (ANN) model for predicting each type of project delay with 20 hidden neurons. The dataset was divided into Training, Validation, and Testing portions, set at 70%, 15%, and 15%, respectively. We selected the Levenberg-Marquardt algorithm as the training method. Also, we tested several configurations by varying the number of the hidden layers from 1 to 6 hidden layers, each one consisting of 20 neurons; no significant improvement was observed. Since the training time of ANN depends on the number of hidden layers, we chose a network with one hidden layer.

As mentioned in Table I, each project has 9 characteristic indices as input variables and three delay types as output variables. The total delay of each project is the summation over y_1 to y_3.

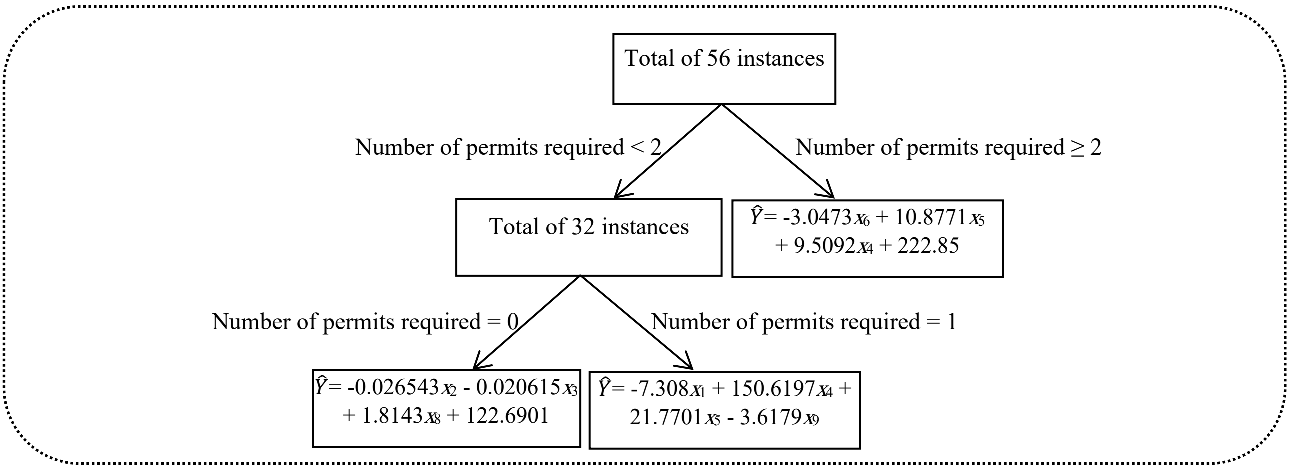
Total of 56 instances
Number of permits required < 2
Number of permits required ≥ 2
Total of 32 instances
$\hat{Y}$= -3.0473x_6 + 10.8771x_5 + 9.5092x_4 + 222.85
Number of permits required = 0
Number of permits required = 1
$\hat{Y}$= -0.026543x_2 - 0.020615x_3 + 1.8143x_8 + 122.6901
$\hat{Y}$= -7.308x_1 + 150.6197x_4 + 21.7701x_5 - 3.6179x_9

Fig. 6. Final view of a Trended RT for predicting total project delay ($Y=y_1+y_2+y_3$).

Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are reported here. They are calculated as follows:

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}_i\right| \tag{11}$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2} \tag{12}$$

As can be seen in Table II, the *k*-NN regression with each of its derivatives mentioned earlier in last section provides a zero error. This zero-error for seen data points is a trivial result of “Zero Distance Exclusive Neighborhood” rule embedded in the *k*-NN regression algorithms. In order to have a more precise judge about the accuracy of the algorithms, their performance should be evaluated for unseen projects.

Table II. Experimental results for predicting all kinds of project delays

Output variable	**Y1**		**Y2**		**Y3**		**Y1+Y2+Y3**	
Algorithm	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
TRT	13.8	17.1	4.8	6.2	5.4	7.2	15.6	18.8
RT	26.1	31.9	6.8	8.2	7.2	9.0	26.8	32.7
k-NN RC	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
k-NN EC – Median	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
k-NN EC – 25-75	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
k-NN EC – 30-70	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ANN	24.7	36.3	6.5	9.0	8.3	10.7	16.7	27.0
SVM	37.3	44.1	6.4	8.3	7.6	10.4	35.6	42.6
CUBIST	14.6	19.0	7.5	9.2	8.0	9.9	16.1	20.5

C. Cross Validation

A common way to evaluate the effectiveness of algorithms is *n*-fold validation. In this approach, the dataset is randomly divided into *n* partitions, and at each of *n* runs of the algorithm, one-fold is labeled as test set and set aside from dataset, while the remaining instances are used to train the algorithm. Finally, the test fold is used for evaluating the algorithm performance when dealt with unseen data points. Table III summarizes the performance of all the algorithms in predicting project delays for unseen projects.

Table III. Experimental Results for predicting all kinds of project delays for unseen data

Output variable	**Y1**		**Y2**		**Y3**		**Y1+Y2+Y3**	
Algorithm	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
TRT	17.7	21.4	8.5	11.0	9.5	12.0	19.9	24.1
RT	27.3	33.1	8.8	10.6	10.0	12.3	27.3	33.0
***k*-NN RC**	49.4	60.9	12.0	14.7	14.9	17.9	49.0	60.9
***k*-NN EC – Median**	22.0	26.7	8.1	9.7	9.6	11.3	19.8	25.2
***k*-NN EC – 25-75**	21.9	26.9	8.2	9.5	9.7	11.3	19.4	25.6
***k*-NN EC – 30-70**	21.7	26.6	8.2	9.6	9.7	11.3	19.7	25.6
ANN	45.0	59.1	12.5	15.9	10.7	13.4	49.4	69.6
SVM	43.3	49.6	7.4	9.2	8.7	11.3	41.5	48.0
CUBIST	21.0	25.2	8.5	10.4	9.6	11.8	20.6	24.5

The best prediction accuracy of each delay type y_i is highlighted in bold. As observed, the Trended Regression Tree provides the best prediction accuracy for Delay of Performance (y_1).

Not only Trended Regression Tree does not yield the best prediction accuracy for estimating Delay of Materials (y_2), and Delay of Inspection (y_3), but also the predictions made with other methods result in drastic relative errors. The root cause is, that these kinds of delays are less predictable and it can be revealed by calculating the relative error of best predictions (Table IV).

As can be inferred from Table IV, relative errors for predicting Delay of Inspection (y_3) and Delay of Materials (y_2) are so high that there will be no profit if we try to predict this kind of project delay even with the best available algorithm (*i.e.* SVM). In Other words, although the SVM reveals the best performance in making predictions, but since it yields an MAE of 7.4, corresponding to a 41% error, it is not rational to recommend SVM as the best method for predicting Delay of Materials in piping projects in Iran.

Table IV. Relative errors for best estimates made through several estimators.

Output variable	**Y1**		**Y2**		**Y3**		**Y1+Y2+Y3**	
Average y_i	128.5		18.0		17.0		163.5	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Best Prediction Error	17.7	21.4	8.1	9.5	9.0	11.3	19.4	24.1
Best Relative Prediction Error (%)	13.8	16.7	45	52.8	52.9	66.5	11.9	14.7

If we try to predict the total project delay ($y_1 + y_2 + y_3$), the forecasts may be more reliable with less relative error. In contrast with the high relative errors associated with predicting the Delay of Inspection (y_3) and the Delay of Materials (y_2), the total project delay as the summation of all delay types yields a lower relative error. Eventually, we propose the Trended Regression Tree and *k*-Nearest Neighbor Regression, along with its two consensus-making versions as median (Eq.6), or 25%-75% percentile (Eq.7) for predicting total project completion delay.

One often-overlooked aspect in the reported research for comparing performance measures of forecasting methods is comparing the quality of forecasting methods based on testing statistical hypotheses. In order to test the strict superiority of the method with the least MAE/RMSE error, several hypothetical tests are available. One of the most specialized tools is the Diebold-Mariano (D-M) statistic (Diebold and Mariano, 1995):

$$D - M = \frac{\bar{d}}{\sqrt{\sigma_d^2 / T}} \tag{13}$$

In which

$$d = e_{i,t}^2 - e_{j,t}^2, \bar{d} = \frac{1}{T}\sum_{t=1}^{T}\left(e_{i,t}^2 - e_{j,t}^2\right) \tag{14}$$

Where $e_{i,t}$, $e_{j,t}$ are the forecast errors of two methods *i* and *j* respectively, and σ_d^2 is a consistent estimator of the asymptotic variance of $\sqrt{T}\bar{d}$. Table V summarizes the hypothetical testing of the superiority of the best estimator over second best estimating algorithm. As it can be inferred from the p-values row in Table V, at a 92% confidence level (*p*-value < 0.08), the TRT outperforms CUBIST in estimating y_1. Another interesting finding of the testing hypothesis is that no other method is better than the second-best algorithm when estimating y_2, y_3, and the summation of all delay types. If we test the superiority of SVM (as the first best estimator) against TRT as the fourth (with respect to MAE) or even seventh (with respect to RMSE) best estimator in predicting y_2, there is no evidence for its superiority at the 90% confidence level.

Table V. Testing the hypothesis of superiority of best estimator over 2d best one.

Output variable	**Y1**		**Y2**		**Y3**		**Y1+Y2+Y3**	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Best Prediction	TRT	TRT	SVM	SVM	SVM	SVM	*k*-NN EC – 25-75	TRT
Best Prediction Error	17.7	21.4	7.4	9.2	8.7	11.3	19.4	24.1
2nd Best Prediction	CUBIST	CUBIST	*k*-NN EC – Median	*k*-NN EC – 25-75	TRT	CUBIST	*k*-NN EC – 30-70	CUBIST
2nd Best Prediction Error	21.0	25.2	8.1	9.5	9.5	11.8	19.7	24.5
D-M statistics	1.4301	1.4301	0.4979	0.2855	0.5925	0.3303	0.1603	0.1416
p-value	0.0760	0.0760	0.309	0.388	0.277	0.371	0.436	0.444

D. Model verification

Based on the performance measures reported in table III, our novel developed algorithms achieve the lowest Mean Absolute Error and Root Mean Squared Error. In a more definite meaning and based on statistical inferences, it can be concluded that TRT outperforms CUBIST, a commercial package for such complex regression problems. On the other hand, two other well-known tools for pattern recognition, such as ANN and SVM, lack the accuracy for predicting the complicated behavior of project delays.

To analyze the effect of key parameters on the performance of all the implemented algorithms, a range of values of those parameters was tested, with a single value from a robust range reported. For example, the number of hidden layers of the ANN was tested up to 20 layers, and no significant change was observed in the performance of ANN.

Throughout the correlation analysis for ignoring dimensions of less correlation with the output variable, a threshold of 65±7 on the percentiles did not make any significant performance difference.

For constructing a binary splitting tree of TRT, the Minimum Parent Size was tested for 30±5, and the results were robust enough to report a value of 30 for that size of the database. The *k* factor for all four *k*-NN regression algorithms was tested for 10±5 and no significant change was observed in the performance of all k-NN derivatives.

VI. MANAGERIAL INSIGHT

Piping projects are vital in Governmental Gas Distribution Companies. Accurate prediction of delays in such projects is essential for contract parties to avoid legal disputes and unfair judgments due to the lack of appropriate foresight. The proposed prediction model is likely to benefit decision-makers by predicting possible delays based on documented factors during project life cycle. The managerial impact of the developed model is expected to pave the way towards broader long-term context for assessing the contractors. Based on the results of this research, we found that the number of permits required for each piping project in Khuzestan is an essential factor determining the delays of each project; so, we advised the management team of Khuzestan Gas Company to seek for the routines to accelerate and facilitate the process of issuing required permits from civil service organizations and local government. Since this factor has the most considerable effect on project delays compared to others, every quest on improving this factor will likely yield immediate and substantial improvements.

VII. SUMMARY AND CONCLUSION

The success of the piping projects in meeting their time and cost goals plays a crucial role in the success of gas distribution companies. Delays can incur additional costs for project stakeholders and make disputes in involved groups, making accurate and clear mechanisms for delay analysis essential to understanding the various effects on parts of projects and revealing the underlying causes of delays.

In this paper, we developed two data mining methods to predict project delays. The main advantage of the prediction models made by Trended Regression Trees over classic function estimation methods is in the managerial interpretations made as post-processing on the fitted model. Conventional classic function estimation methods such as ANN, SVM regression, or other regression methods act as a black box to model the output variable. In contrast, regression trees effectively recognize and clearly visualize the main effective factors in shaping the current state of the system.

For example, our analysis revealed that if the number of required permits exceeds one permit through the overall progress of the project, then a delay type categorized as performance delay will have a nearly threefold sudden amplification. This kind of clear project delay modeling will enable the company to make serious investigations into the root causes of such drastic performance defeats. Experimental results also reveal a significant performance superiority of our developed algorithm over previous decision tree-based algorithms, such as classic Regression Tree and CUBIST software.

Improving the quality of forecasts in every field of management is highly advantageous, especially in managing construction and piping projects. Usually, one of the main contests in piping projects between contractors and Governmental Gas Distribution Companies (GGDC) is the root cause of delays. If the contractor is capable of justifying that the delays arise from the GGDC, no penalty is charged; otherwise, a penalty is charged relative to the delay period. Using the findings of this research help the GGDCs to be capable of judging the root cause of delays based on the last judged root causes of project delays. Another side benefit of the findings of this research is preparing managerial insights for the GGDCs to analyze the severity of the effects of each kind of delay factor. This can result in reducing the magnitude of factors or taking precautionary actions by the contractor or the GGDC.

Being more informed about the possible project delays provides a competitive advantage to both contractors and GGDCs in form of project resiliency. Specifically, resilience in projects is defined as the capability of a project to respond to, prepare for, and reduce the impact of disruptions caused by the drifting environment and project complexity. Since accurate forecasting of project delays could prepare managers with enough respite to plan for successful reactive tasks. In this research, we developed two strong tools to make accurate predictions. Since these tools are derived from machine learning literature, they are capable of being tailored for pattern recognition in other fields of management, such as predicting price fluctuations and demand forecasts.

REFERENCES

Adam, A., Josephson, P., & Lindahl, G. (2017). Aggregation of factors causing cost overruns and time delays in large public construction projects: Trends and implications. *Engineering, Construction and Architectural Management, 24*, 393-406.

Al-Kharashi, A., & Skitmore, M. (2009). Causes of delays in Saudi Arabian public sector construction projects. *Construction Management and Economics, 27*, 3-23.

Alshboul, O., Alzubaidi, M.A., Mamlook, M.E., Almasabha, G., Almuflih, A.S., & Shehadeh, A. (2022a). Forecasting Liquidated Damages via Machine Learning-Based Modified Regression Models for Highway Construction Projects. *Sustainability*, 14(10), 5835.

Alshboul, O., Shehadeh, A., Al Mamlook, R.E., Almasabha, G., Almuflih, A.S., & Alghamdi, S.Y. (2022). Prediction Liquidated Damages via Ensemble Machine Learning Model: Towards Sustainable Highway Construction Projects. *Sustainability, 14(15)*, 9303.

Alshboul, O., Shehadeh, A., Almasabha, G., & Almuflih, A.S. (2022b). Extreme Gradient Boosting-Based Machine Learning Approach for Green Building Cost Prediction. *Sustainability*, 14(11), 6651.

Borovsky, A., Thal, D., & Leonard, L.B. (2021). Moving towards accurate and early prediction of language delay with network science and machine learning approaches. *Scientific Reports* (11), 8136.

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199-231.

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (1983). Classification and Regression Trees, Taylor & Francis.

Chan, D.W., & Kumaraswamy, M.M. (2002). Compressing construction durations: lessons learned from Hong Kong building projects. *International Journal of Project Management*, 20, 23-35.

Banerjee Chattapadhyay, D., Putta, J., & Rao P, R. M. (2021). Risk identification, assessments, and prediction for mega construction projects: A risk prediction paradigm based on cross analytical-machine learning model. *Buildings*, *11*(4), 172.

Cozad, A., Sahinidis, N.V., & Miller, D.C. (2014). Learning surrogate models for simulation-based optimization. *Aiche Journal, 60*(6), 2211-2227.

Cui, Y., Liu, H., Wang, Q., Zheng, Z., Wang, H., Yue, Z., ... & Yao, M. (2022). Investigation on the ignition delay prediction model of multi-component surrogates based on back propagation (BP) neural network. *Combustion and Flame*, *237*, 111852.

Davoudabadi, R., Mousavi, S.M., Šaparauskas, J., & Gitinavard, H. (2019). Solving construction project selection problem by a new uncertain weighting and ranking based on compromise solution with linear assignment approach. *Journal of Civil Engineering and Management*, 25(3), 241-251.

Derakhshanfar, H., Ochoa, J. J., Kirytopoulos, K., Mayer, W., & Langston, C. (2020). A cartography of delay risks in the Australian construction industry: impact, correlations and timing. *Engineering, Construction and Architectural Management*, 28(7), 1952–1978.

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, *20*(1), 134-144.

Doraisamy, S. V., Akasah, Z. A., & Yunus, R. (2015). An overview on the issue of delay in the construction industry. In *InCIEC 2014: Proceedings of the International Civil and Infrastructure Engineering Conference 2014* (pp. 313-319). Springer Singapore.

Durdyev, S., & Hosseini, M.R. (2020). Causes of delays on construction projects: a comprehensive list. *International Journal of Managing Projects in Business*, 13(1), 20-46.

Egwim, C.N., Alaka, H., Toriola-Coker, L.O., Balogun, H., & Sunmola, F. (2021). Applied artificial intelligence for predicting construction projects delay. *Machine Learning with Applications*, Volume 6, 100166.

Abd El-Razek, M. E., Bassioni, H. A., & Mobarak, A. M. (2008). Causes of delay in building construction projects in Egypt. *Journal of construction engineering and management*, *134*(11), 831-841.

Fallahnejad, M. (2013). Delay causes in Iran gas pipeline projects. *International Journal of Project Management, 31(1)*, 136-146.

Friedman, J.H. (1991). Multivariate Adaptive Regression Splines. *Annals of Statistics, 19*(1), 1-67.

Ghazal, M.M., & Hammad, A. (2022). Application of knowledge discovery in database (KDD) techniques in cost overrun of construction projects. *International Journal of Construction Management*, *22(9)*, 1632-1646.

Gitinavard, H. (2019). Strategic evaluation of sustainable projects based on hybrid group decision analysis with incomplete information. *Journal of Quality Engineering and Production Optimization*, 4(2), 17-30.

Gitinavard, H., & Mousavi, S. M. (2015). Evaluating construction projects by a new group decision-making model based on intuitionistic fuzzy logic concepts. *International Journal of Engineering*, *28*(9), 1312-1319.

Gitinavard, H., Mousavi, S., Vahdani, B., & Siadat, A. (2020). Project safety evaluation by a new soft computing approach-based last aggregation hesitant fuzzy complex proportional assessment in construction industry. *Scientia Iranica*, 27(2), 983-1000.

Gunduz, M., Nielsen, Y., & Ozdemir, M. (2015). Fuzzy Assessment Model to Estimate the Probability of Delay in Turkish Construction Projects. *Journal of Management in Engineering, 31(4)*, 04014055.

Gurgun, A.P., Koc, K., & Kunkcu, H. (2022). Exploring the adoption of technology against delays in construction projects, Engineering, *Construction and Architectural Management*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/ECAM-06-2022-0566

Hamzeh, A. M., Mousavi, S. M., & Gitinavard, H. (2020). Imprecise earned duration model for time evaluation of construction projects with risk considerations. *Automation in Construction*, 111, 102993.

Ilic, I., Görgülü, B., Cevik, M., & Baydogan, M.G. (2021). Explainable boosted linear regression for time series forecasting. *Pattern Recognition, 120*, 108144.

Islam, M.S., & Trigunarsyah, B. (2017). Construction Delays in Developing Countries: A Review. *Journal of Construction Engineering and Project Management, 7(1)*, 1-12.

A Kassem, M., Khoiry, M. A., & Hamzah, N. (2021). Theoretical review on critical risk factors in oil and gas construction projects in Yemen. *Engineering, Construction and Architectural Management*, *28*(4), 934-968.

Kleijnen, J.P. (2017). Regression and Kriging metamodels with their experimental designs in simulation: A review. *European Journal of Operational Research, 256*, 1-16.

Klumpenhouwer, W., & Shalaby, A. (2022). Using Delay Logs and Machine Learning to Support Passenger Railway Operations. Journal of the Transportation Research Board, 2676(9). https://doi.org/10.1177/03611981221085

Korhonen, K.T., & Kangas, A.S. (1997). Application of nearest-neighbour regression for generalizing sample tree information. *Scandinavian Journal of Forest Research, 12*, 97-101.

Li, M., Vanberkel, P., & Zhong, X. (2022). Predicting ambulance offload delay using a hybrid decision tree model. *Socio-Economic Planning Sciences*, *80*, 101146.

Lin, C., & Fan, C. (2019). Evaluation of CART, CHAID, and QUEST algorithms: a case study of construction defects in Taiwan. *Journal of Asian Architecture and Building Engineering, 18*, 539 - 553.

Loh, W. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1*. 14–23.

Mahmoodzadeh, A., Nejati, H.R., & Mohammadi, M. (2022). Optimized machine learning modelling for predicting the construction cost and duration of tunneling projects. *Automation in Construction*, Volume 139, 104305.

Mehrabi Sharafabadi, H., & Movafaghpour, M.A. (2021). Investigating Causes of Delay in Natural Gas Distribution Pipeline Projects: a Correlation Analysis (Case Study: Khuzestan Province of Iran). *Journal of Applied Research on Industrial Engineering, 9(1)*, 68-77.

Mittas, N., & Mitropoulos, A. (2022). A Data-Driven Framework for Probabilistic Estimates in Oil and Gas Project Cost Management: A Benchmark Experiment on Natural Gas Pipeline Projects. *Computation*, 10(5), 75. https://doi.org/10.3390/computation10050075

Mohammed, R.M., & Suliman, S.M. (2019). Delay in Pipeline Construction Projects in the Oil and Gas Industry: Part 1 (Risk Mapping of Delay Factors). *International Journal of Construction Engineering and Management, 8(1)*, 24-35.

National Iranian Gas Company Website, https://nigc.ir/index.aspx?siteid= 1&&site ID=1&pageid=172

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical *Computing*. Vienna, Austria.

RuleQuest (2016). Data mining with cubist. https://www.rulequest.com/cubist-info.html.

Sambasivan, M., Deepak, T.J., Salim, A., & Ponniah, V. (2017). Analysis of delays in Tanzanian construction industry: Transaction cost economics (TCE) and structural equation modeling (SEM) approach, *Engineering, Construction and Architectural Management*, 24 (2), 308-325.

Sanni-Anibire, M.O., Zin, R.M. & Olatunji, S.O. (2022). Machine learning model for delay risk assessment in tall building projects. *International Journal of Construction Management*, Volume 22(11).

Seber, G., & Lee, A. (2012). Linear regression analysis. *Wiley Series in Probability and Statistics*. Wiley.

Sen, A., & Srivastava, M. (2012). *Regression analysis: Theory, methods, and applications*. Springer New York.

Shoar, S., Chileshe, N., & Edwards, J.D. (2022). Machine learning-aided engineering services' cost overruns prediction in high-rise residential building projects: Application of random forest regression. *Journal of Building Engineering*, Volume 50, 104102.

Smola, A., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14(3)*, 199-222.

Taleongpong, P., Hu, S., Jiang, Z., Wu, C., Popo-Ola, S., & Han, K. (2021). Machine learning techniques to predict reactionary delays and other associated key performance indicators on British railway network. *Journal of Intelligent Transportation Systems, 26(3)*, 311-329. https://doi.org/10.1080/15472450.2020.1858822

Türkakin, O. H., Manisali, E., & Arditi, D. (2020). Delay analysis in construction projects with no updated work schedules. Engineering, *Construction and Architectural Management, 27(10)*, 2893–2909.

Yang, J. B., & Wei, P. R. (2010). Causes of delay in the planning and design phases for construction projects. *Journal of Architectural Engineering*, *16*(2), 80-83.

Yang, L., Liu, S., Tsoka, S., & Papageorgiou, L.G. (2017). A regression tree approach using mathematical programming. *Expert Systems With Applications 78*, 347–357.

Zhang, N., & Wei, G. (2013). Extension of VIKOR method for decision making problem based on hesitant fuzzy set, *Applied Mathematical Modelling, 37(7)*, 4938-4947.

Zhang, Y., & Sahinidis, N.V. (2013). Uncertainty Quantification in CO2 Sequestration Using Surrogate Models from Polynomial Chaos Expansion. *Industrial & Engineering Chemistry Research, 52(9)*, 3121-3132.