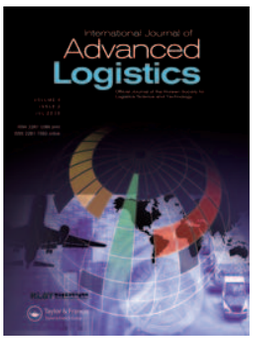


International Journal of Advanced Logistics

ISSN: 2287-108X (Print) 2287-7592 (Online) Journal homepage: http://www.tandfonline.com/loi/tjal20

A machine learning approach for predicting delays in construction logistics

Ahmad Asadi, Mohammed Alsubaey & Charalampos Makatsoris

To cite this article: Ahmad Asadi, Mohammed Alsubaey & Charalampos Makatsoris (2015) A machine learning approach for predicting delays in construction logistics, International Journal of Advanced Logistics, 4:2, 115-130, DOI: 10.1080/2287108X.2015.1059920

To link to this article: http://dx.doi.org/10.1080/2287108X.2015.1059920

Published online: 01 Sep 2015.

International Journal of Advanced Logistics, 2015
Vol. 4, No. 2, 115–130, http://dx.doi.org/10.1080/2287108X.2015.1059920

Taylor & Francis
Taylor & Francis Group

A machine learning approach for predicting delays in construction logistics

Ahmad Asadi, Mohammed Alsubaey and Charalampos Makatsoris*

Department of Mechanical, Aerospace and Civil Engineering, College of Engineering and Physical Sciences, Brunel University London, London, UK

Construction project management is vital for accomplishing pre-determined objectives. Despite using construction management, most of the projects do not meet original time schedule or has been delayed. Delay is one of the biggest problems faced by construction industry. This project is a study the critical delay factors for project management in construction focusing contractors in Qatar and to build a prediction model to avoid the same in future projects. The objectives of this research project are to investigate delay factors to help contractors to reach their goals on time during construction. This research will review the delay factors through literature review survey questionnaire targeting professionals at a construction company who are involved in many construction projects in Qatar. The correlation between them is examined to produce the best ways in preventing delays.

This study was carried out based on comprehensive literature review, which was done to provide the background, history and delay factors of delays in construction. The information of literature review was then used to design and conduct a survey questionnaire to investigate delay factors in construction projects in Qatar and was distributed to the targeted respondents at the contractors company. Later the top delay factors achieved from the questionnaire were combined with secondary data collected from an ongoing mega project for the same company to build a prediction model using WEKA software.

Keywords: delay factors; construction industry; project management; delay prediction; data mining; machine learning; Qatar

1. Introduction

A construction project is commonly acknowledged as successful, when it is completed on time, within budget, in accordance with the specifications and to the stakeholders' satisfaction. Functionality, profitability to contractors, absence of claims and court proceeding and "fitness for purpose" for occupiers have also been used as measures of project success.

One of the most important problems in the construction projects is delay. Delays occur in every construction project and the magnitude of these delays varies considerably from project to project. Some projects are only a few days behind the schedule: some are delayed over a year. So it is essential to define the actual causes of delays in order to minimize and avoid the delays in any construction projects.

Delay is a situation when the contractor, consultant, and client jointly or individually contribute to the non-completion of the project within the original, specified or agreed contract period. Delays causes disruption of work and loss of productivity, late completion of project, increased time-related cost, and third-party claims, and abandonment or termination of contract. It is important that general management keep track of project progress to minimize the possibility of delay occurrence or identify it at early stages.

The construction industry is one of the industries that involved many uncertainties in its everyday operations. The study of recent literature shows that construction projects are normally accomplished with large cost overruns, extended schedules (delays), and quality concerns.

A study in Hong Kong [6] reviewed the causes of construction delays as seen by clients, consultants, and contractors, and then examined the factors affecting productivity. The study exposed differences in perceptions of the relative significance of factors between the three groups, indicative of their experiences, possible prejudices and lack of effective communication. The impact of delays may include time overrun, cost overrun, disputes, arbitration, litigation, and total abandonment.

Other studies have been carried out to investigate the factors that lead to successful completion of projects. Some researcher evaluates the concept of success in a construction project when the evaluation dimensions are adequately

*Corresponding author. Email: harris.makatsoris@brunel.ac.uk

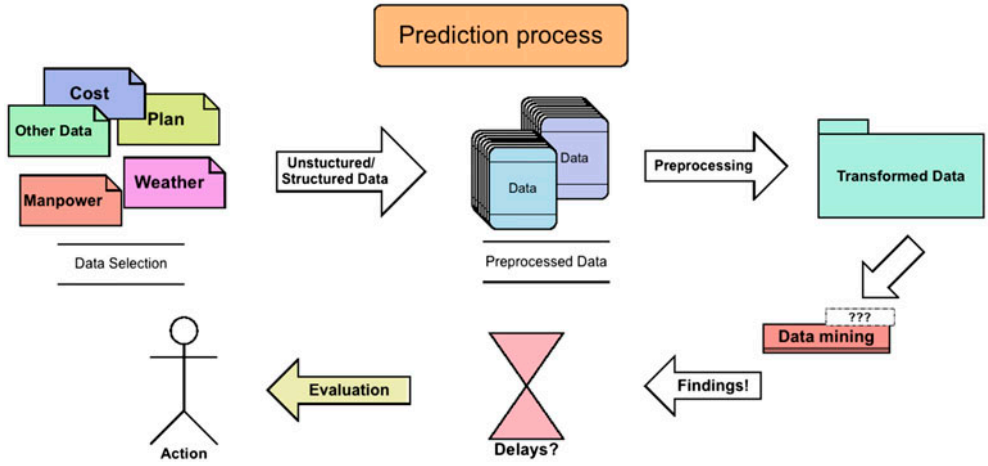


Figure 1. Delay prediction process for this study.

defined. The first study to determine critical success factors was carried out [4]; the researcher identified which factors were most important in successfully completing construction projects. Another researcher [19] indicates that research into critical success factors has been undertaken since 1967, and demonstrates the development of information on critical success factors based on empirical and theoretical studies [25].

The construction projects success in Qatar is currently low. Therefore, this paper focuses on:

1. What are the real causes of project delays in the Qatari Construction industry?
2. How to predict delays for similar projects with the use of data mining?

A survey was designed to answer these questions and a data-mining prediction model was tested on WEKA software to improve construction industry performance for similar projects (Figure 1).

Delay has significant effect on completion cost and time of construction project. Delays can be minimized when their causes are identified. Knowing the cause of any particular delay in a construction project would help avoiding the same.

Although various studies have been undertaken to identify the factors affecting the causes of delays, since the problems are rather contextual, the studies need to focus on specific geographical area, country, or region. A major criticism of the Qatari construction industry is due to the growing rate of delays in projects delivery.

2. Delays in construction projects

Despite a wealth of research into the causes of delays in projects, time and cost remains a continuing problem in the construction industry while the delay prediction research in this field is limited. In this study, the literature on delays in construction was revised with a view to use data-mining methods to predict delays and avoid them of accruing in future projects.

2.1. Common causes for delays

Given the impacts of project delays causes, it has been broadly studied by researchers from different countries [1–3, 5, 6, 11, 15, 18, 20–24, 28]. Table 1 summarizes the general causes of project delays, which were conducted in a survey and identified in this study.

3. Research methodology

The methodologies and procedures employed for the study includes data collection, sampling (populations used), questionnaire design, the data analysis, and then combining the data to use it in WEKA software for classification and

prediction. Four types of methodological approaches were employed in this study. They are the quantitative, qualitative, triangulation, and historical approaches.

3.1. Research design and data collection process

3.1.1. Survey data

The survey was designed based on the review of relevant literatures and questionnaire surveys, 49 triggers of delay were identified as shown in previous Table 1, to which participants were asked to indicate the level of importance of each cause. These causes were categorized into three main sets:

Table 1. Common causes of delays.

1	Regular interference and poor communication	Client
2	Implementation of safety procedures	
3	Variation order and late approval for payment	
4	Late supply of information and late decision-making	
5	Project objectives are not very clear	
6	Nomination of Subcontractors and suppliers	
7	Many provisional sums and prime cost	
8	Duration is not enough for constructing the project	
9	Irregular payments and disturbed cash flow of main contractor	
10	Routine of government authorities and approvals	
11	Irregular attending of weekly meetings	
12	Incomplete contract documents	Consultant
13	Incomplete drawings	
14	Poor design management	
15	Slow response	
16	Delayed approval of drawings and quantities for construction	
17	Inadequate duration for inspection	
18	Experience of staff in management and technical inspection	
19	Delay in submittal and approval	
20	Poor-quality control	
21	Poor communication between consultant staff	
22	Incompetence to contractor's technical enquiries	
23	Changes in drawings and specifications	
24	Incompetent technical staff	
25	Shortage of materials or Delay of delivery	
26	Inappropriate organization management	Contractor
27	Implementation of safety procedures	
28	Lack of technical professional in the organization	
29	Unsmooth external and internal communications	
30	Lack coordination with subcontractors	
31	Centralization with top management	
32	Delayed mobilization	
33	Incompetent contractor staff	
34	Poor planning, scheduling or resource management	
35	Poor-quality control	
36	Congested construction site	
37	Mistakes during construction	
38	Lack of experience of similar projects	
39	Materials specifications	
40	Shortage of materials	
41	Delay of delivery	
42	Poor quality of materials	
43	Shortage of manpower	
44	Low productivity	
45	Unavailability of equipments	
46	Unexpected weather conditions	
47	Different nationalities of work force	
48	Preparing the method statement for each activity	
49	Work permits	

- **The Client:** Related factors to Clients include finance, payments, Safety procedures, client interference, slow decision, and unrealistic contract duration.
- **The Consultant:** Related factors Consultants include contract management, preparation and approval of submittals, and materials and site supervision.
- **The Contractors:** Related factors to Contractors include site management, improper planning, Safety procedures, inadequate contractor experience, mistakes during construction and preparing the construction methodology.

Other significant issues of concerns such as project location, site conditions, neighbors, and changes, materials including quality and shortage and the nominated Sub-contractors were also considered in the questionnaire survey in Qatar.

Questionnaire was distributed to professionals at the Contractors Company, who were working on mega, large, and medium projects in the State of Qatar with project values from US $ 50 Million to US $ 400 Million. The survey was conducted by project managers/directors, construction managers, project engineers, section and site engineers who have specific experience in this area that can provide an insight into both underlying causes of delay and their relative effect. The response rate of questionnaire was 100%.

3.1.2 Project reports

Real data from project reports such as monthly progress reports and manpower plan were gathered for 24 months as shown below Table 2. In order to combine them with other delay factors in a later stage to prepare the training set which will be used in WEKA tool for classification and prediction purpose.

4. Data analysis

4.1. Analysis of survey data

The results of the survey were analyzed using the Rating Scale method (Wuensch, 2009) also known as Likert scale [17]. Rating Scale questions calculate a weighted average based on the weight assigned to each answer choice [7] the rating average is calculated as follow:

w =weight of answer choice, x =response count for answer choice

$$\frac{x_1w_1 + x_2w_2 + x_3w_3 \ldots x_nw_n}{Total}$$

4.2. Data preparation

Nowadays there is huge amount of data being collected and stored in databases everywhere across the globe. The tendency is to keep increasing year after year. It is not hard to find databases with Terabytes of data in enterprises and research facilities. There is invaluable information and knowledge hidden in such databases; and without automatic methods for extracting this information it is practically impossible to mine for them. For example, this study will highlight how this hidden data can be used with the real data such as project reports to predict delays in construction projects.

Table 2. Project reports data.

Month	Monthly plan	Cumulative plan	Manpower plan	Manpower actual	Monthly actual	Actual cumulative
Oct-11	0.08	0.08	2000	1750	0.06	0.06
Nov-11	0.14	0.22	3100	2775	0.13	0.19
Dec-11	0.29	0.51	4200	3850	0.17	0.36
Jan-12	0.7	1.21	4800	4555	0.9	1.26
Feb-12	1.11	2.32	5400	5100	0.79	2.05

The "Unexpected weather conditions" delay factor was used with the real-project report data as a sample in this study to predict delays in construction projects, based on the survey results since it was selected as one of the top factors that could cause delays to construction projects in the state of Qatar. It is monitored on daily basis by the safety department to calculate the heat index and to check if the site conditions are suitable for workers. However, these data then are stored and never used for prediction purposes.

Throughout the years many algorithms were created to extract what is called nuggets of knowledge from large sets of data. There are several different methodologies to approach this problem: classification, association rules, clustering, etc. This study will focus on *classification*.

4.3. Knowledge extraction algorithms and their implementation

WEKA [27] contains modules for data pre-processing, classification, clustering, and association rule extraction.

In this study, data pre-processing and two classification algorithms are used in order to predict delays established on the available data.

4.3.1. Classification algorithm

Classification is a data mining (machine learning) technique used to predict a certain outcome for data instances based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm attempts to discover relationships between the attributes that would make it possible to predict the outcome. Next the algorithm is given a data-set not seen before, called prediction set, which contains the same set of attributes, except for the prediction attribute. The algorithm analyses the input and produces a prediction. The prediction accuracy defines how suitable the algorithm is. In this study, the construction database is the training set which has relevant project information recorded previously, where the prediction attribute is whether or not the progress had a delay. Tables 2 and 3 below illustrate the training and test sets of such database.

In this study, two classifiers, "J48 decision tree" algorithm and "Naive Bayes" algorithm are used for comparison. Comparison is made on accuracy, sensitivity, and specificity using true positive and false positive in confusion matrix generated by the respective algorithms. Also we can use the correct and incorrect instances that give us a most efficient method for classification by using the confusion matrix [16].

1. J48 Decision tree classifier. Is supervised learning. The algorithm J48 is an extension of the C4.5 classifier. A decision tree is a tree-like structure, where each internal node denotes a test on a data attribute, each branch represents an outcome of the test, and each leaf node holds a class label. The topmost node in a tree is the root node. During tree construction, attribute selection measures like information gain, gain ratio, gain index are used to select the attribute that best partitions the tuples into distinct classes. Model generated by decision tree helps to predict new instances of data [10].

While building a tree, J48 ignores the missing values, i.e. the value for that item can be predicted based on what is known about the attribute values for the other records. The basic idea is to divide the data into range based on the attribute values for items that are found in the training sample. J48 allows classification by either decision trees or rules generated from them [26].

2. Naive Bayes classifier. Given a set of data, each of which belongs to a known class, and has a known vector of variables, this study is to build a rule that will allow us to assign future objects to a class, given only the vectors of variables describing the future objects. Problems of this kind are called supervised classification; several methods for building such rules have been developed. Naive Bayes, independence Bayes and simple Bayes [13], is a simple probabilistic classifier, based on the Bayesian theorem with strong (naive) independence assumptions. It can predict class

Table 3. Training set.

Month	Monthly plan	Man-power plan	Man-power actual	Monthly cost in M	Out-look	Humidity	Temp	Wind	Delay
May-12	1.49	8000	7625	10	Sunny	25	35	False	No
Jun-12	1.8	8400	8200	10	Sunny	35	38	False	Yes
Jul-12	3.34	9100	8755	10	Hot	60	40	False	Yes

membership probabilities, such as the probability that a given tuple belongs to a particular class. It uses prior probability of each category given no information about an item. Categorization produces a posterior probability distribution over the possible categories given a description of an item [12]. The more general version of Bayes rule formula can be written as:

$$P(C|A_1, A_2 \ldots A_n) = \frac{(\prod_{i=1}^{n} P(A_i|C))P(C)}{P(A_1, A_2 \ldots A_n)}$$

Where, C is a class value, and the attributes are A1, A2…An.

The Naive Bayes model is one of the oldest formal classification algorithms, and even in its simplest form it is often surprisingly effective. It is widely used in areas such as text classification and spam filtering. The statistical, data mining, machine learning, and pattern recognition communities have introduced a large number of modifications in an attempt to make it more flexible, but such modifications are necessarily complications, which detract from its basic simplicity [9].

4.3.2. *Model building*

4.3.2.1. Data Preparation. To identify the classification accuracy of the previous techniques, a training set was provided and cleaned (Figure 2) by removing invalid data and supplying them with missing values to make sure that a reliable result is obtained. The training set is actually the historical data of an ongoing Oil and Gas mega-project in the state of Qatar combined with weather conditions in the same area. The data-set included many attributes; such as Monthly plan, cumulative plan, manpower plan and actual, monthly cost (according to *Middle East Economic Digest*), cumulative cost, the weather conditions; and the delay, which identifies whether there is a delay in progress or not.

The data were stored in Excel and saved as Comma Separated Value (CSV) format. The CSV file was imported in Notepad and was converted into an Attribute Relation File Format (ARFF) file. An ARFF file has three components: @relation which gives the name of the data-set, @attribute which identifies the elements of the tables with the corresponding value, and @data which lists all the records. With total of 34 instances in the training set only.

```
@relation training_set
@attribute “monthly plan” numeric
@attribute “cumulative plan” numeric
>
@attribute windy {FALSE,TRUE}
@attribute delay {No,Yes}
@data
0.08,0.08,2000,1750,10,10,Sunny,29,25,FALSE,No
0.14,0.22,3100,2775,10,20,Rainy,20,18,TRUE,No
>
```

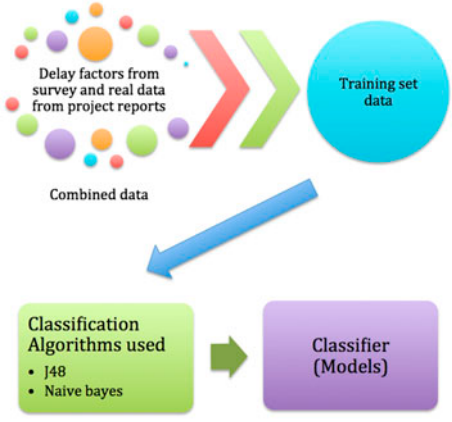


Figure 2. Classification process: model construction.

4.96,90.97,9554,9226,10,340,Hot,60,38,FALSE,Yes
2.64,93.61,9289,8450,10,350,Hot,70,40,FALSE,Yes

4.3.2.2. Input data into WEKA. Once the ARFF training file is loaded, the histogram can be displayed and visualized for all attributes as shown in below Figure 3, the attributes pane shows the variables from the selected training data-set that has 11 attributes including monthly plan, manpower plan, cost, weather conditions, and delays (progress).

The first classifiers J48 followed by Naive Bayes were then selected from the Classify tab for the test option. Once the first model is created, it validates the accuracy result of the model for each classifier.

As there are satisfactory amount of data, the data are divided into three different data-sets: training set, validation set, and test set [8], where the training set is used to train different models. Two test files (validation/test) were tested on the classifier with higher accuracy using the option-supplied test set as illustrated below (Figure 4).

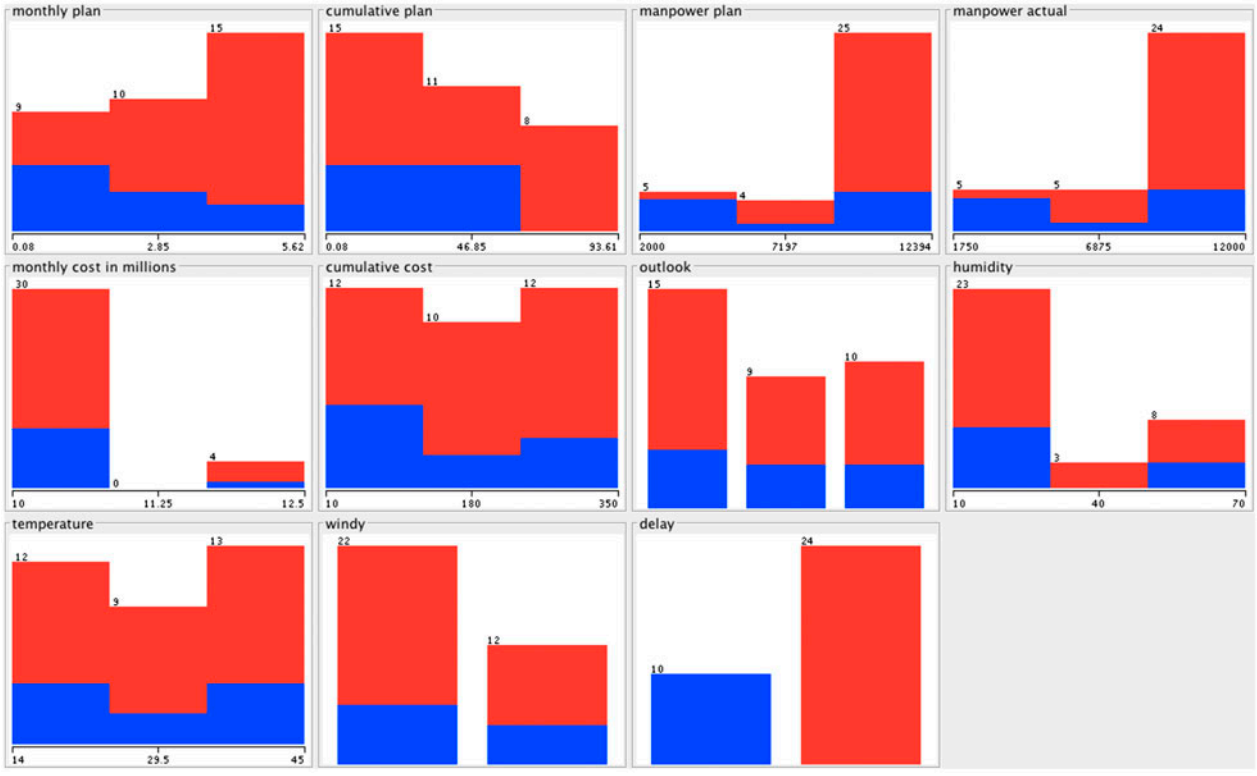


Figure 3. A visual representation of all the attributes selected.

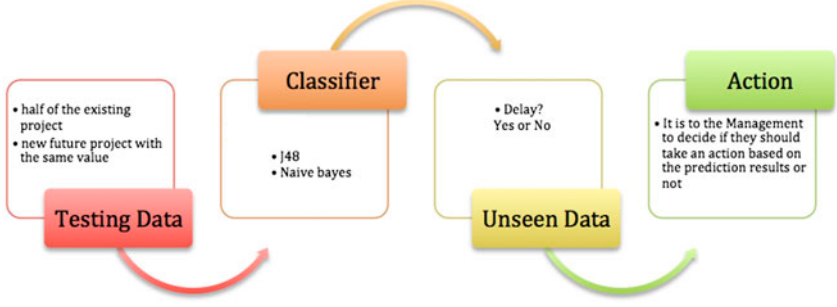


Figure 4. Testing process: Use the model in Prediction.

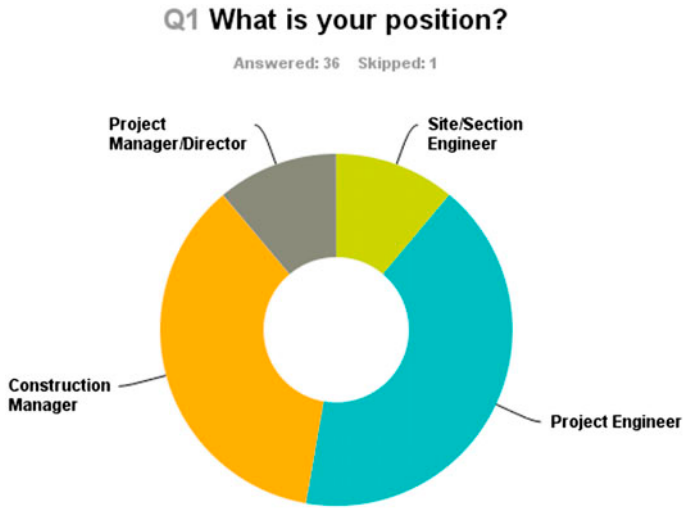


Figure 5. Respondents' position.

4.3.2.2.1. Test set for the existing project (validation). This is a validation set, which is a sample of 16 instances (Aprit 2013 to July 2014, 16 months) taken from the existing project, which was loaded as training set, to check if similar readings will appear.

4.3.2.2.2. Test set for new project with the same value. For this test set, 34 instances were used but the records were not included in the training set data, in order to get delay prediction results for new future project that starts in October 2015 with the similar value of period, cost, manpower, and different weather conditions. Results and functionality of the models are discussed below.

5. Results and discussion

The first section describes the results and discussion from the survey and the second section shows the results and discussion for the prediction model using WEKA software.

5.1. Survey results

Thirty-seven professionals working at a Construction Company in Qatar filled the survey (Figure 5).

The result of the analysis question 3,4 and 5 shows that:

- The client's involvement in causes of delay is 13.04%
  (3 out of 23 major causes)
- The consultant's involvement in causes of delay is 34.78%
  (8 out of 23 major causes)
- The contractor's involvement in causes of delay is 52.17%
  (12 out of 23 major causes)

The result also demonstrates a high consistency between the causes of delay in Qatar construction projects (Figures 6, 7, 8) and the general causes identified by previous researchers.

Results from the delays factors related to Clients, shows that Routine government authorities and approvals ranked with the highest total average rating value of 4.19, while late supply of information and late decision-making is ranked second with total average rating value of 4.06. There is a close interrelation among the factors that were ranked the highest, where the government authorities (approvals from civil defense, for example) can be involved in changing the design of any project that will lead the clients to late decision-making. Duration is not enough for constructing the project and implementation of safety procedures were ranked the least with values of 3.34 and 3.25 respectively.

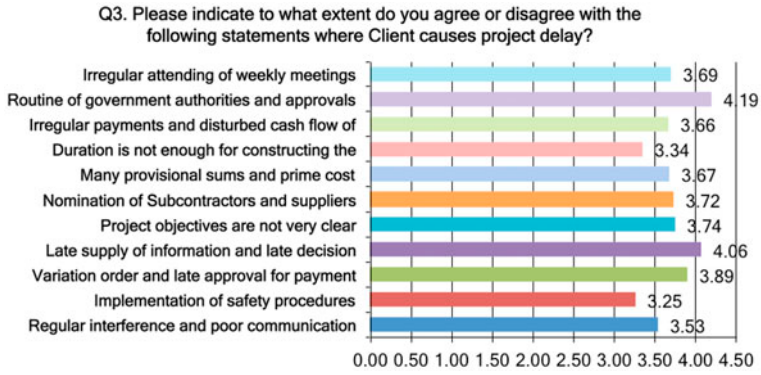


Figure 6. Delay Factors caused by the clients.

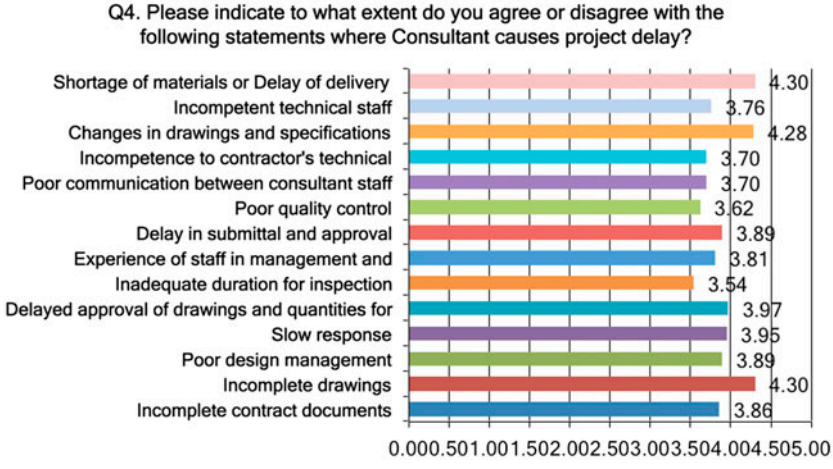


Figure 7. Delay Factors caused by the consultants.

From below figure, and based on the total average rating for consultants, it is found that the most frequent causes of delay as ranked by the respondents are the ones related to materials and drawings with highest value of 4.30 for both factors the shortage of material or delay in delivery and the incomplete drawings from designers, followed by changes in drawing and specifications with value of 4.28 and the delayed approval of drawings and quantities of material at 3.97. The relation between the top factors is noticed; large margin of errors will appear in the drawings prepared by the consultants and this could be due to the pressure they face from clients or the conflicts that may appear for drawings from different departments such as instruments and piping.

The inadequate duration for inspection and poor-quality control factors were ranked the lowest factors that may cause delays with values of 3.54 and 3.62, respectively. This is suitable because of the strict rules that apply to the quality assurance departments to make sure the projects meet the standards of quality and it is carried out by qualified inspectors.

For the last question, the survey shows that the respondents feel the shortage of material is an important delay factor, which affects the contractors; it was ranked at total average rating value of 4.62. Followed by the lack of technical professional ranked at value of 4.57 and delay of delivery at 4.51. The study also shows the importance of other factors that can cause delays to the contractor, inappropriate organization management factor was ranked at value of 4.46, followed by PTW at 4.27 and unexpected weather conditions at 4.19.

The lowest score of 2.97 was given to the different nationalities of workforce factor; it is not a concern if workers are assigned to a work supervisor who speaks the same language for the ease of communication.

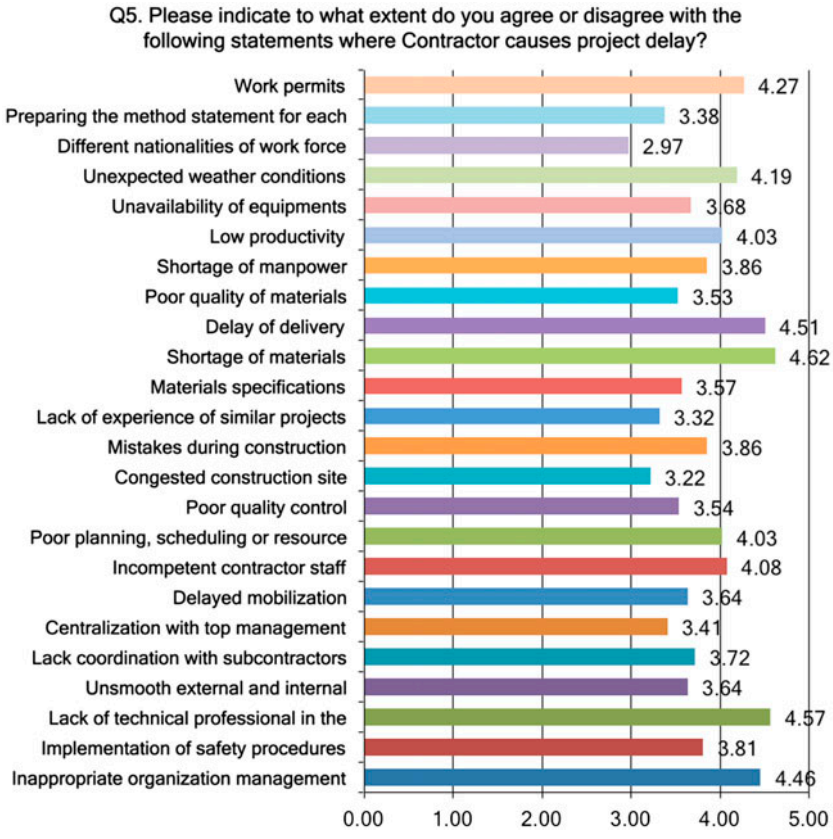


Figure 8. Delay Factors caused by the contractors.

5.2. *Delay predictions*

5.2.1. *Experiment and results*

After pre-processing the data on WEKA and applying the classification techniques twice on the training set. Each classifier generates rules and displays output to all parameters such as confusion matrix as shown in below figures. It took both classifiers 0.1 s to build the classifier model. The test option uses Cross Validation by default with 10 folds, this means that the data-set is split into 10 parts, the first 9 are used to train the algorithm, and the 10th is used to assess the algorithm. This process is repeated allowing each of the 10 parts of the split data-set a chance to be the held out test set.

The simulations results are then compared and segregated into several sub items for the ease of analysis. On the first part correctly and incorrectly classified instances (Prediction Accuracy) are shown in numeric and percentage values followed by the Kappa statistics (Table 5).

Furthermore, Table 6 shows the result based on error during the simulation. Mean absolute error and root mean squared errors are in numeric values.

Below figures 9 and 10 shows the graphical representation of the results.

Table 4. Validation or test sets.

Month	Monthly plan	Man-power plan	Man-power actual	Monthly cost in M	Out-look	Humidity	Temp	Wind	Delay
Jan-14	4.47	9840	9720	15	Rainy	10	14	TRUE	?
Feb-14	5	9840	9666	15	Sunny	12	20	TRUE	?
Mar-14	4.63	9840	9450	15	Sunny	18	25	FALSE	?

Table 5. Simulation result for each algorithm.

Algorithm	Correctly Classified Instances %	Incorrectly Classified Instances %	Kappa statistic
J48	79.41% (27)	20.58% (7)	0.377
NaiveBayes	73.52	26.47	0.343

Table 6. Training and simulation errors.

Algorithm	Mean Absolute error	Root Mean Squared Error	Relative Absolute Error %	Root Relative Squared Error %
J48	0.2939	0.4219	69.7837	92.4672
NaiveBayes	0.3026	0.496	71.8621	108.7054

Table 7. Prediction accuracy.

Month	Actual progress (Delay)	Predicted results (Delay)	Remarks
Apr-13	Yes	Yes	TRUE
May-13	Yes	Yes	TRUE
Jun-13	Yes	Yes	TRUE
Jul-13	No	Yes	FALSE
Aug-13	No	Yes	FALSE
Sep-13	No	Yes	FALSE
Oct-13	No	Yes	FALSE
Nov-13	Yes	Yes	TRUE
Dec-13	Yes	Yes	TRUE
Jan-14	Yes	Yes	TRUE
Feb-14	Yes	Yes	TRUE
Mar-14	Yes	Yes	TRUE
Apr-14	Yes	Yes	TRUE
May-14	Yes	Yes	TRUE
Jun-14	Yes	Yes	TRUE
Jul-14	Yes	Yes	TRUE

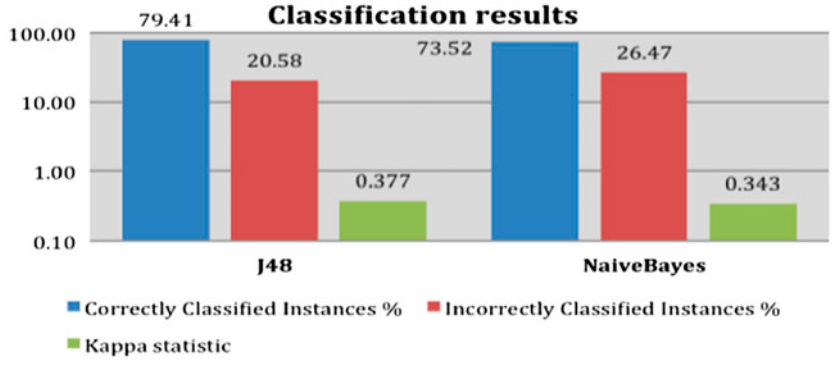


Figure 9. Results.

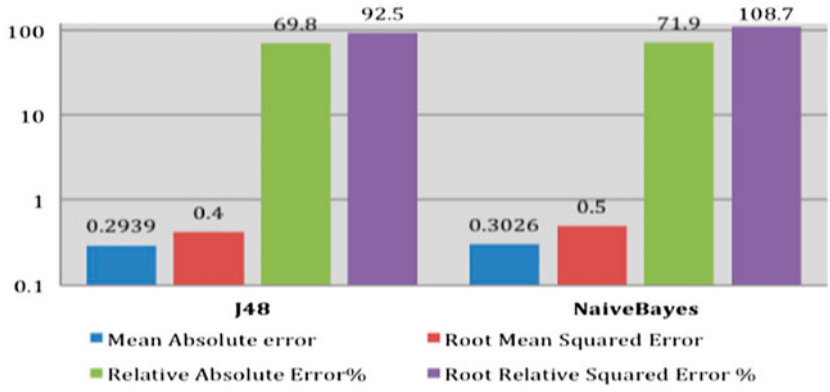


Figure 10. Error rate of classifiers.

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.3       0          1         0.3       0.462      0.556    No
                 1         0.7        0.774     1         0.873      0.556    Yes
Weighted Avg.    0.794     0.494      0.841     0.794     0.752      0.556

=== Confusion Matrix ===

  a  b   <-- classified as
  3  7 |  a = No
  0 24 |  b = Yes
```

Figure 11. Confusion matrix generated by J48.

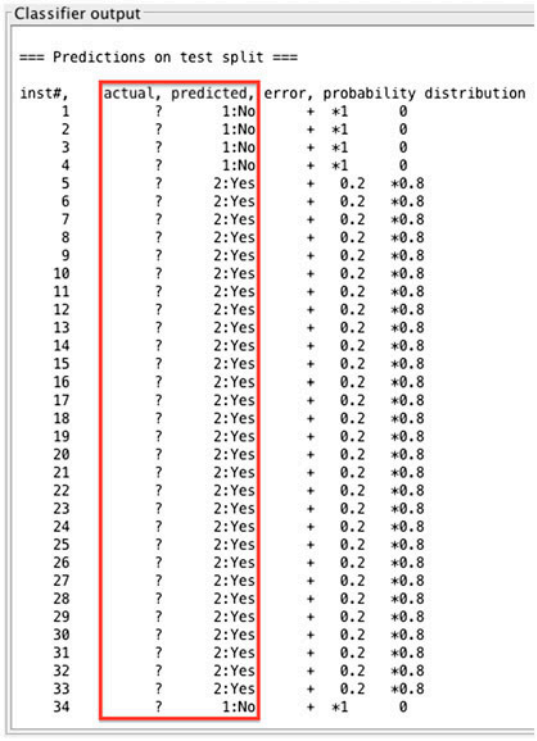
```
Classifier output

Time taken to build model: 0 seconds

=== Predictions on test split ===

inst#,     actual, predicted, error, probability distribution
     1        ?      2:Yes      +   0.2   *0.8
     2        ?      2:Yes      +   0.2   *0.8
     3        ?      2:Yes      +   0.2   *0.8
     4        ?      2:Yes      +   0.2   *0.8
     5        ?      2:Yes      +   0.2   *0.8
     6        ?      2:Yes      +   0.2   *0.8
     7        ?      2:Yes      +   0.2   *0.8
     8        ?      2:Yes      +   0.2   *0.8
     9        ?      2:Yes      +   0.2   *0.8
    10        ?      2:Yes      +   0.2   *0.8
    11        ?      2:Yes      +   0.2   *0.8
    12        ?      2:Yes      +   0.2   *0.8
    13        ?      2:Yes      +   0.2   *0.8
    14        ?      2:Yes      +   0.2   *0.8
    15        ?      2:Yes      +   0.2   *0.8
    16        ?      2:Yes      +   0.2   *0.8
```

Figure 12. Prediction results for existing project.

Based on the results from tables and figures, it can be seen clearly that there is not much difference in the readings but still the J48 algorithm gives better result and it can perform better; it has higher accuracy 79.41% compare to Naive Bayes algorithm which showed a lower accuracy value at 73.52%.

According to the confusion matrix results for J48 the test sets were executed on J48 classifier model for prediction as it has better accuracy results.

Confusion matrix is generated for class delay having two possible values, i.e. YES or NO. (Figure 11).



Classifier output

=== Predictions on test split ===

inst#	actual	predicted	error	probability distribution
1	?	1:No	+	*1 0
2	?	1:No	+	*1 0
3	?	1:No	+	*1 0
4	?	1:No	+	*1 0
5	?	2:Yes	+	0.2 *0.8
6	?	2:Yes	+	0.2 *0.8
7	?	2:Yes	+	0.2 *0.8
8	?	2:Yes	+	0.2 *0.8
9	?	2:Yes	+	0.2 *0.8
10	?	2:Yes	+	0.2 *0.8
11	?	2:Yes	+	0.2 *0.8
12	?	2:Yes	+	0.2 *0.8
13	?	2:Yes	+	0.2 *0.8
14	?	2:Yes	+	0.2 *0.8
15	?	2:Yes	+	0.2 *0.8
16	?	2:Yes	+	0.2 *0.8
17	?	2:Yes	+	0.2 *0.8
18	?	2:Yes	+	0.2 *0.8
19	?	2:Yes	+	0.2 *0.8
20	?	2:Yes	+	0.2 *0.8
21	?	2:Yes	+	0.2 *0.8
22	?	2:Yes	+	0.2 *0.8
23	?	2:Yes	+	0.2 *0.8
24	?	2:Yes	+	0.2 *0.8
25	?	2:Yes	+	0.2 *0.8
26	?	2:Yes	+	0.2 *0.8
27	?	2:Yes	+	0.2 *0.8
28	?	2:Yes	+	0.2 *0.8
29	?	2:Yes	+	0.2 *0.8
30	?	2:Yes	+	0.2 *0.8
31	?	2:Yes	+	0.2 *0.8
32	?	2:Yes	+	0.2 *0.8
33	?	2:Yes	+	0.2 *0.8
34	?	1:No	+	*1 0

Figure 13. Prediction results for new project.

And this is a specific table layout that allows visualization of the performance of an algorithm. For above highlighted confusion matrix, true positive for class a = No is 3 while false positives is 7. Whereas, for class b=Yes, true positives is 24 and false positives is 0, i.e. 24+3 = 27 which represents the correct instances classified and the other elements = 7+0 = 0 represents the incorrect instances. [10]

True positive rate = diagonal element/ sum of relevant row

False positive rate = non-diagonal element/ sum of relevant row

Precision = diagonal element/ sum of relevant column

Hence,

TP Rate for class a = 3/(3+7) = 0.3

FP Rate for class a = 0/(0+24) = 0

TP Rate for class b = 24/(24+0) = 1

FP Rate for class b = 7/(7+3) = 0.7

Precision for class a = 3/(3+0) = 1

Precision for class b = 24/(24+7) = 0.774

5.3. Discussion

The test sets were loaded on the J48 classifier model developed in earlier stage since it revealed better accuracy readings; the prediction results for both sets are presented below.

5.3.1. Prediction for existing project (Validation model)

This step was completed for validation of the model and to check prediction accuracy since the data used in this test set is already included in the training set, 16 instances (From Apr-13 to Jul-14) and their progress was already identified in terms of delays from the real data collected. The predicted results showed that there would be delays for the 16 months (instances) as shown in Figure 12.

The actual unknown delay values are identified as **"?"**, where the predicted values are specified as "Yes" in above output. A comparison between the achieved prediction results from the validation set and the real data monitored in the project reports is shown in Table 7.

We can note that the predicted results matched the original project report by 75% accuracy with 12 true values out of 16. The accuracy percentage is also very close to 79.4%, which is the correctly classified number of instances given by J48 classifier.

If carried out properly, and if the validation set and training set are from the same population, validation is nearly fair. However, if it is misused, the prediction errors in the validation are likely to be much worse than would be expected.

5.3.2. *Prediction for new project*

Care should be taken while selecting the test sets. Especially when the test set is representative of the training set. According to [14], *"validation only yields meaningful results if the validation set and training set are drawn from the same population."* If a model is developed to predict a risk of delays from occurring in a new project, the value of the new project used in the test set for prediction should be similar to the existing project. If the model is trained using data from a study involving only a specific population group (e.g. Medium project), but is then applied to the general population (e.g. Mega project for Oil and Gas), the results from the training set could differ greatly from the actual predictive performance. Therefore, this test set is a prediction for a new project of the same value as the existing one in terms of duration, cost and manpower; the data used were not included in the training set.

Prediction results achieved for new project are as shown in Figure 13.

The results show that the new project will be completed on time and has no delays in the first four months. However, delays will start to occur from the fifth month onwards to almost the end of the scheduled plan.

This prediction results for the future projects will help project managers to identify the causes of delays that will occur from the fifth month of the scheduled plan and will allow them to change the project resources and costing to the best of their plan and to reduce risks of delay.

Predictable delays can also be factored into the project via a documented risk management plan. When that plan is prepared, risks can be identified and evaluated to determine probable delays, and potential mitigating action. If the predicted delays do come to pass, the risk management plan will provide a pre-planned course of action.

6. Conclusions and future work

Delays are a major problem faced by most of the construction projects in Qatar while the causes of delays and the actions to avoid them are still not fully understood by the project participants. This research discussed the specific operational environment of Qatar construction projects and investigated the causes of the delay contributed by the project participants.

Unlike other countries or regions, the construction projects in Qatar face many challenges such as cultural, high-quality and architecture requirements, shortage of workforce, and hot weather conditions. These constraints make the causes of delays different from other countries. The consultants play a very important role in design and drawings-related delays because they are in charge of the design process in conjunction with the owner of the project. On the other hand, the contractor has the major responsibility for delays in construction-related delays and material procurement process.

This study has explored different models and various classification methods, interesting results has been found on predictability of project delays by measuring the performance of the models using real data. The paper also presented new project delay prediction model based on data mining. The results obtained for two classification algorithms showed an average test accuracy of 76% when classifying delays from a combination of real-project monitored data and weather conditions delay factor, for 34 months duration.

The best algorithm for the data used from construction project is J48 classifier with accuracy of 79.41% and the total time taken to build the model is at 0.1 s. Naïve Bayes classifier has lowest accuracy and higher error rate compared to J48. Despite the small differences in readings obtained for the two classifiers, these results suggest that among the machine-learning algorithm tested, J48 classifier has the potential to significantly improve the conventional classification methods for use in construction field.

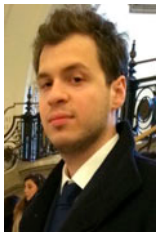
Data mining offers promising ways to expose hidden patterns within a large amount of data, which can be used to predict future behaviors . The results achieved for project delay predictions would help managers to categorize risks associated with execution stages allowing them to identify delays before they happen and their main causes rather than minimize the consequences as well as measure the progress to success. Establishing predictions and thus assumptions engage managers to chase specific information and more importantly define their value, which in turn help to determine the areas of concerns in which the managers should be willing to improve.

This paper had mixed of success in attaining the required goals to investigate and identify delay factors affecting construction projects in the state of Qatar, and to build a model to predict whether a future project will be delayed or not, by evaluating two classification algorithms based on WEKA software.

To the best of our knowledge the proposed models present the first attempt to predict construction delays using WEKA and the models could be improved in several ways including trying other classification techniques that may produce a better model or by adding more data in the training set in order to get higher accuracy rate, this could be done by collecting data from daily progress report instead of months to increase the number of instances. Similarly the prediction model could be refined, in order to give specific predictions for activities with certain durations, such as cable laying or concrete pouring plan. The model can also be used to predict safety-related incidents for specific activities such incidents related to hot season. We can imagine building a phenomenon model, which will use a combination of different models to select the best one for different cases to predict delays for new projects without the need of several months of data that can be completed by building separate models for specific durations or area of project and then grouping them into a common model.

Finally, the most interesting area would be progress and delay prediction model for construction projects using data mining for image processing, this could be done by building a tool to compare real-time images taken from site to 3D design models for the same location.

Notes on contributors



Ahmad Asadi has extensive engineering experience and possesses excellent management skills. Prior to completing his MSc degree in Engineering Management at Brunel University of London; Ahmad graduated with a BEngg (Honors) Degree in Electrical and Electronic Engineering from London and thereafter worked with an international company (CCIC) on some major construction projects involving large companies in Qatar. His current research interests include applications of machine learning and data mining in project delay prediction. Ahmad is working toward Chartered Engineer status and he is a member of the Institute of Engineering and Technology of UK and the Institute of Electrical and Electronics Engineers.

Mohammed Alsubaey is currently a PhD candidate in Enterprise System Engineering at Brunel University of London; his research focus is on construction project early warnings using unstructured data by Machine Learning Logic. He earned his MSc degree in Engineering Management from Brunel University in 2011. Mohammed also holds a Bachelor degree in Architecture and building technology from KFU University of Dammam, Saudi Arabia (2009). Prior to his PhD candidacy he worked as Project Control Engineer on some mega projects in Saudi Arabia and has a broad range of engineering and management experiences to date.

Charalampos Makatsoris is a Chartered Engineer and Reader in Manufacturing and Engineering Systems in the Department of Mechanical, Aerospace and Civil Engineering at Brunel University London. He holds a first degree and PhD in Manufacturing Engineering from Imperial College London and has spent nearly 20 years in both industry and academia. He leads a multi-disciplinary team undertaking research in manufacturing process design, optimization, and automation. He has authored over 70 papers to date and one book on Manufacturing ICT published by Springer (Kluwer). He has long-standing expertise in manufacturing systems engineering. Specifically he is interested on artificial intelligence and machine learning in engineering design and manufacture. His research has attracted significant funding to date from several funding bodies and involves several academic and industrial partners. He also has a track record in program management and technology transfer having been involved in University spin out companies from their conception through to successful implementation of exit strategy. He is currently member of the board of directors on two overseas companies both in the area of ICT.

ORCID

Charalampos Makatsoris http://orcid.org/0000-0003-0139-6791

References

[1] Abd Majid MZA, McCaffer R. Factors of non-excusable delays that influence contractors' performance. J Manage Eng. 1998;14 (3):42–49.

[2] Aibinu AA, Jagboro GO. The effects of construction delays on project delivery in Nigerian construction industry. Int J Project Manage. 2002;20:5939.
[3] Alkass S, Mazerolle M, Harris F. Construction delay analysis techniques. Construction Manage Econ. 1996;14:375–394.
[4] Ashley DB, Lurie CS, Jaskelskis EJ. Determinants of construction project success. Project Manage J. 1987;18:69–79.
[5] Assaf SA, Al-Khalil M, Al-Hazmi M. Causes of delays in large building construction projects. J Constr Eng Manage, ASCE. 1995;11(2):45–50.
[6] Chan DWM, Kumaraswamy MM. A comparative construction durations: lessons learned from Hong Kong building projects. Int J Project Manage. 2002;20:22–35.
[7] Dawes J. Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. Int J Market Res. 2008;50(1):61–77.
[8] Dmitrienko A, Chuang-Stein C. Pharmaceutical statistics using SAS: A practical guide. 2nd ed. USA: SAS Institute Inc.; 2007. p. 46–53.
[9] Domingos P, Pazzani M. On the optimality of the simple Bayesian classifier under zero-one loss. Mach Learn. 1997;29:103–130.
[10] Dunham, MHH. Data mining introductory and advanced topics. Upper Saddle River, N.J.: Prentice Hall. chapter 2.4; 2003.
[11] Frimpong Y, Oluwoye J, Crawford L. Causes of delay and cost overruns in construction of groundwater projects in a developing countries; Ghana as a case study. Int J Project Manage. 2003;21:321–326.
[12] Han. J, Kamber. M, Pei. J. Data mining: Concepts and techniques. 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 2011.
[13] Hand DJ, Yu K. Idiot's Bayes – not so stupid after all? Int Stat Rev. 2001;69:385–398.
[14] Hirsch R. Validation Samples. Biometrics. 1991;47(3):1193–1194.
[15] Koushki PA, Al-Rashid K, Kartam N. Delays and cost increases in the construction of private residential projects in Kuwait. J Constr Manage Econ. 2005;23(3):285–294.
[16] Lee W, Stolfo SJ, Mok KW. A data mining framework for building intrusion detection models. In: Proceedings of the 1999 IEEE Symposium on Security and Privacy, 1999. 1999 May 9-12; Oakland, CA, USA.
[17] Likert R. A technique for the measurement of attitudes. Arch Psychology. 1932;140:1–55.
[18] Mansfield NR, Ugwuo O, Doran T. Causes of delay and cost overruns in Nigerian construction projects. J Constr Eng Manage. 1994;13(4):254–260.
[19] Mengesha WJJ. Performances for Public Construction Projects in Developing Countries: Federal Road & Educational Building Projects in Ethiophia Norwegian University of Science and Technology. Doctoral Thesis 2004:45.
[20] Mezher TM, Tawil W. Causes of delays in the construction industry in Lebanon. Eng Constr Architectural Manage. 1998;5 (3):252–260.
[21] Odeh AM, Battaineh HT. Causes of construction delay: traditional contracts. Int J Project Manage. 2002;20:67–73.
[22] Odeyinka H, Yusif A. The causes and effect of construction delays of housing projects in Nigeria. J Finance Manage Property Constr. 1997;3(2):31–44.
[23] Ogunlana SO Prokuntong K, Jearkjirm, V. Construction delays in fast growing economy comparing Thailand with other economies. Int J Project Manage. 1996;14(1):37–45.
[24] Ren Z, Atout M, Jones J. Root causes of construction project delays in Dubai. In: Dainty A, editor. Procs 24th Annual ARCOM Conference, 1–3 September 2008. Association of Researchers in Construction Management: Cardiff, UK; 2008. p. 749–757.
[25] Ruben IM, Seeling W. Experience as a factor in the selection and performance of project managers. IEEE Trans Eng Manage. 1967;EM-14(3):131–135.
[26] Spangler W, May J, Vargas, L. Choosing data-mining methods for multiple classification: Representational and performance measurement implications for decision support. J Manage Inf Syst. 1999;16(1):37–62.
[27] WEKA. Website: Machine learning group. Weka 3: Data Mining Software in Java. Available: http://www.cs.waikato.ac.nz/ml/weka/. Last accessed Sep 2014.
[28] Wiguna IPA, Scott S. Nature of the critical risk factors affecting project performance in Indonesian building contracts. In: Khosrowshahi F (Ed.), 21st Annual ARCOM Conference, 7-9 September 2005, SOAS, University of London. Association of Researchers in Construction Management, Vol. 1; 2005. p. 225–235.