Machine Learning Algorithms for Construction Projects Delay Risk Prediction

Ahmed Gondia, S.M.ASCE[1]; Ahmad Siam[2]; Wael El-Dakhakhni, F.ASCE[3]; and Ayman H. Nassar[4]

Abstract: Projects delays are among the most pressing challenges faced by the construction sector attributed to the sector's complexity and its inherent delay risk sources' interdependence. Machine learning offers an ideal set of techniques capable of tackling such complex systems; however, adopting such techniques within the construction sector remains at an early stage. The goal of this study was to identify and develop machine learning models in order to facilitate accurate project delay risk analysis and prediction using objective data sources. As such, relevant delay risk sources and factors were first identified, and a multivariate data set of previous projects' time performance and delay-inducing risk sources was then compiled. Subsequently, the complexity and interdependence of the system was uncovered through an exploratory data analysis. Accordingly, two suitable machine learning models, utilizing decision tree and naïve Bayesian classification algorithms, were identified and trained using the data set for predicting project delay extents. Finally, the predictive performances of both models were evaluated through cross validation tests, and the models were further compared using machine-learning-relevant performance indices. The evaluation results indicated that the naïve Bayesian model provides a better predictive performance for the data set examined. Ultimately, the work presented herein harnesses the power of machine learning to facilitate evidence-based decision making, while inherent risk factors are active, interdependent, and dynamic, thus empowering proactive project risk management strategies. **DOI: 10.1061/(ASCE)CO.1943-7862.0001736.** *© 2019 American Society of Civil Engineers.*

Author keywords: Classification; Complex systems; Confusion matrices; Construction projects; Cross validation; Delay risk analysis; Machine learning; Predictive data analytics; Risk identification; Time delay.

Introduction

Construction project delay is a global phenomenon (Assaf and Al-Hejji 2006; Sambasivan and Soon 2007). Its occurrence is mainly attributed to the interdependent inherent risk factors and uncertainties associated with the complex and dynamic nature of construction processes. Such factors might be related, for example, to stakeholder(s) incompetence, poor communication, inadequate estimation of employed resources, contractual deviations, or even municipal constraints (Assaf and Al-Hejji 2006; Aziz 2013). As a result, project delays, quantified through time overrun (TO), can negatively impact the project and its stakeholders in multiple ways, including: (1) claims, disputes, and arbitration; (2) cost overruns and loss of revenue; (3) disruption of work and loss of productivity; or (4) contract termination and, possibly, total project abandonment (Aibinu and Jagboro 2002; Majid 2006).

In order to provide accurate estimates of project durations, construction firms typically adopt standard quantitative delay risk analysis tools. For example, Monte Carlo analysis can be used for investigating the complete extent of risk associated with scheduled work items to estimate more reasonable project completion dates (Rezaie et al. 2007; Sadeghi et al. 2010; Kokkaew and Wipulanusat 2014). Similar to most probabilistic modeling tools, Monte Carlo analysis is data-intensive and requires estimates of work item duration ranges and relevant probability distributions. These estimates can be acquired either *objectively*, through historical data of similar projects, or, more commonly in the absence of such data, *subjectively* based on expert opinion and judgement. The latter approach, however, poses several key limitations related to: (1) using imprecise and/or ambiguous data sources that would typically add another layer of uncertainty to the analysis; (2) overlooking the complex and interdependent nature of inherent risk factors in construction projects that might significantly influence original predictions; and (3) using data that are rarely updated to represent the actual project progressions since the subjective data collected are relevant only at the time of apprehension (Ferson 2008; Guyonnet et al. 2003; Goldstein 2006; Tixier et al. 2017). To overcome these limitations, it is critical to base construction schedule estimates and planning decisions on knowledge extracted from objective and factual data.

In recent times, the construction sector has experienced an explosive growth in the amount of such objective data generated on a daily basis and stored from the various disciplines throughout the project or the facility lifespan (Bilal et al. 2016). Such provision of data creates an opportunity for extracting useful corporate knowledge and promising solutions to the prevailing project delay dilemma. However, because of the interdependence of construction-related data, the adoption of effective data analytics tools is key.

[1] Ph.D. Candidate, Dept. of Civil Engineering, McMaster Univ., 1280 Main St. West, Hamilton, ON, Canada L8S 4L7 (corresponding author). ORCID: https://orcid.org/0000-0001-6584-2514. Email: gondiaa@mcmaster.ca

[2] Postdoctoral Fellow, Dept. of Civil Engineering, McMaster Univ., 1280 Main St. West, Hamilton, ON, Canada L8S 4L7. Email: siamas@mcmaster.ca

[3] Director, The INViSiONLab, Dept. of Civil Engineering, McMaster Univ., 1280 Main St. West, Hamilton, ON, Canada L8S 4L7. Email: eldak@mcmaster.ca

[4] Assistant Professor in Construction and Project Management, Dept. of Civil Engineering, German Univ. in Cairo, Al Tagamoa Al Khames, Cairo 11835, Egypt. ORCID: https://orcid.org/0000-0001-7421-3496. Email: ayman.hamdy@guc.edu.eg

Note. This manuscript was submitted on September 3, 2018; approved on May 13, 2019; published online on October 28, 2019. Discussion period open until March 28, 2020; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Construction Engineering and Management***, © ASCE, ISSN 0733-9364.

In this respect, the potential of machine learning (ML) techniques and algorithms in analyzing voluminous, complex, and interdependent data sets of varying structures for deriving useful insights cannot be overemphasized (Kim et al. 2008; Bilal et al. 2016). ML algorithms, in their different forms, have been widely used in different fields over the past two decades. Nonetheless, ML remains a new prospect within the construction sector despite its highly regarded advantageous potential. A literature survey by the authors of the present study showed that only a limited number of studies have focused on applications of ML techniques within construction research in general and, to an even much lesser extent, on delay risk analysis. A noncomprehensive list of ML techniques applied in construction-related disciplines includes artificial neural networks (Elazouni 2006; Chao and Chien 2009; Heravi and Eslamdoost 2015), decision trees (Desai and Joshi 2010; Chi et al. 2012; Chou and Lin 2013), logistic regression (Cheung et al. 2010; Alzahrani and Emsley 2013; Hwang and Kim 2016), naïve Bayesian models (Jiang and Mahadevan 2008; Gong et al. 2011; Gerassis et al. 2017), and support vector machines (Cheng and Wu 2009; Lam et al. 2010; Huang and Tserng 2018).

The *goal* of the present study was to identify and develop an efficient predictive data analytics tool to analyze and learn from objective delay risk sources based on previous construction project data. Achieving this goal will ultimately facilitate more accurate predictions of future project durations based on these projects' inherent and expected risk levels, thus supporting a proactive project risk management strategy.

In fulfillment of the stated research goal, the present study was focused on achieving two key objectives (Fig. 1). The *first objective* was to identify relevant delay risk sources and factors extracted from literature and adapted by the industry, and subsequently compile and understand a relevant historical construction project delay risk data set, as similar data are currently not available in open literature to the best of the authors' knowledge. Within the process of data collection, constraints pertaining to data characteristics, project types, and analysis limitations were set. These constraints and limitations ensured the consistency and homogeneity of the input data to the predictive data analytics tool in order to realize meaningful results. Afterward, and from various unstructured historical construction project data formats, data pertaining to different types of delay risk sources, along with their level of contribution to project delay, were extracted and subsequently preprocessed to constitute a structured, consistent, and multivariate data set ideally suited for predictive analytics. Subsequently, an exploratory data analysis was conducted to explore the data set properties and the complex interdependencies between the risk sources was uncovered.

Based on the complexity and interdependence of construction delay risk sources, a ML approach was deemed the most appropriate to tackle such a challenging system of interacting variables, and the rationale behind this selection was reported. As such, the *second objective* of the present study was to identify, develop, and validate appropriate ML algorithms to analyze the compiled previous project data set. Appreciating the data set size and properties and prior to conducting the ML-based analysis, a review of previous ML applications was conducted in order to identify the ML technique and algorithms best suited to the data set considered. Subsequently, the study focused on applying a supervised learning classification technique through two ML algorithms: decision tree and naïve Bayesian classification algorithms, which were found to be ideal considering the compiled data set properties. These two algorithms were used to analyze the degrees of variabilities of the influencing risk sources (the independent variables) and their effect on the extents of TO (the dependent variable) described as class labels, in order to generate two TO predictive models/classifiers. Finally, validation of the two generated models' predictive performances was conducted, in terms of both training and testing, through comparisons of predicted and actual class outcomes; and the two models were evaluated and compared using confusion matrices and multiple performance indices.

Fig. 1 shows the methodology adopted to attain the study's objectives, and thus its goal, described previously. The following sections explain the different steps outlined in Fig. 1 followed by concluding remarks, inferred findings, and future recommendations to reach the research long-term goals.

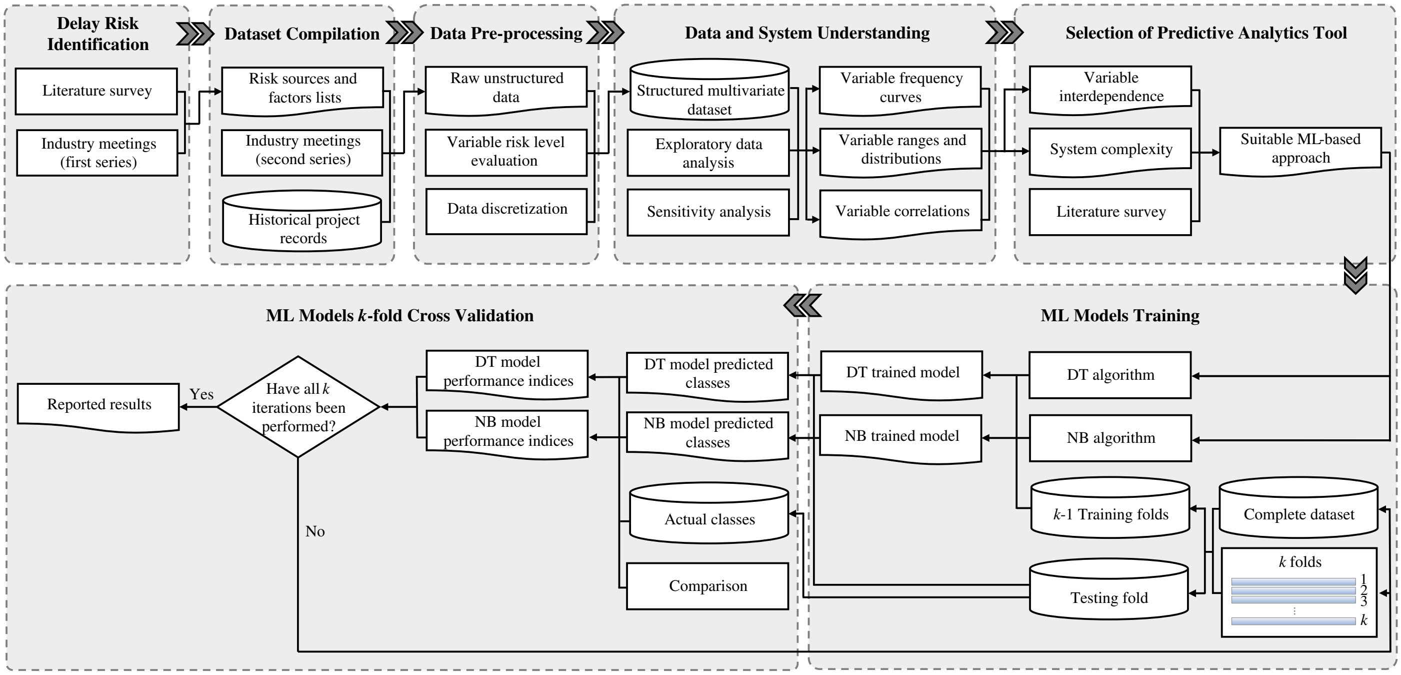

Delay Risk Identification
Literature survey
Industry meetings (first series)
Dataset Compilation
Risk sources and factors lists
Industry meetings (second series)
Historical project records
Data Pre-processing
Raw unstructured data
Variable risk level evaluation
Data discretization
Data and System Understanding
Structured multivariate dataset
Exploratory data analysis
Sensitivity analysis
Variable frequency curves
Variable ranges and distributions
Variable correlations
Selection of Predictive Analytics Tool
Variable interdependence
System complexity
Literature survey
Suitable ML-based approach
ML Models k-fold Cross Validation
Reported results
Yes
Have all k iterations been performed?
No
DT model performance indices
NB model performance indices
DT model predicted classes
NB model predicted classes
Actual classes
Comparison
ML Models Training
DT trained model
NB trained model
DT algorithm
NB algorithm
k-1 Training folds
Complete dataset
Testing fold
k folds
1
2
3
⋮
k

Fig. 1. Analysis methodology and objectives.

Table 1. Previous studies from which risk sources and comprising risk factors were identified

Risk source	Relevant study
1. Owner	Arditi et al. (2006), Chan and Kumaraswamy (1997), Mezher and Tawil (1998), Al Momani (2000), Odeh and Battaineh (2002), Assaf and Al-Hejji (2006), Sambasivan and Soon (2007), and Fugar and Agyakwah-Baah (2010)
2. Consultant	Al Momani (2000), Odeh and Battaineh (2002), Assaf and Al-Hejji (2006), and Aziz (2013)
3. Contractor	Arditi et al. (2006), Chan and Kumaraswamy (1997), Mezher and Tawil (1998), Sambasivan and Soon (2007), and Fugar and Agyakwah-Baah (2010)
4. Design	Arditi et al. (2006), Chan and Kumaraswamy (1997), Assaf and Al-Hejji (2006), and Aziz (2013)
5. Labor	Al Momani (2000), Assaf and Al-Hejji (2006), and Sambasivan and Soon (2007)
6. Materials	Mansfield et al. (1994), Al Momani (2000), Sambasivan and Soon (2007), and Fugar and Agyakwah-Baah (2010)
7. Equipment	Mansfield et al. (1994), Assaf and Al-Hejji (2006), and Sambasivan and Soon (2007)
8. Project	Mansfield et al. (1994), Chan and Kumaraswamy (1997), Fugar and Agyakwah-Baah (2010), and Aziz (2013)
9. External	Chan and Kumaraswamy (1997), Al Momani (2000), and Assaf and Al-Hejji (2006)

Delay Risk Factor and Source Identification

Prior to identifying the type of data to be collected, a literature survey was first conducted in order to identify the most common risk factors influencing building construction project delay and to group such factors into different source categories. This literature survey was conducted in three tiers.

First, a broad search was carried out to identify articles containing the following terms: ("delay" OR "time delay" OR "time overrun" OR "delay risk" OR "delay factor" OR "delay source" OR "delay cause") AND ("construction" OR "construction project"). This search was conducted through two main sources: (1) academic literature databases, including the American Society of Civil Engineers Library, Elsevier Science Direct Digital Library, Springer, Taylor & Francis Online, and Emerald Insight; and (2) academic literature search engines as Web of Science, EBSCOhost, and Google Scholar. The temporal range of the search was set to cover the period from 1980 to 2018, since construction project delay and the factors influencing it have been attracting increased attention for the past three to four decades. By the end of the first tier of literature survey, a total of 83 articles were identified.

In the second tier, the search was narrowed down, whereas the titles, abstracts, and keywords of the articles identified from the first tier were reviewed to select and retain those articles of relevance to the research scope for a full review. Specifically, the following criteria were set for selecting an article for a full review: (1) peer-reviewed articles published in refereed journals of project management, construction management, built environment, or construction economics; and (2) articles focusing on the causes of delays in building construction projects and/or the quantitative assessment of these causes on influencing project delays. This screening process resulted in the selection of 34 relevant research articles from the following journals: *Construction Management and Economics*, *International Journal of Construction Management*, *Journal of Construction Engineering and Management*, *Journal of Management in Engineering*, *International Journal of Project Management*, *Automation in Construction*, and *Construction Economics and Building*.

The third tier involved an in-depth review of the 34 selected articles from the second tier to examine previous studies' identifications of individual delay risk factors and their assemblies into main delay risk sources. Different authors focused on selected risk sources within their articles and identified lists of individual risk factors within such sources. After reviewing the 34 articles, the results of 10 articles were used for risk factor and source identification since the former were repeatedly cited by the rest of the 34 articles. Beyond these 10 articles, no distinct risk factors were identified within any risk source. By the end of the three-tier literature survey, nine delay risk sources were established and subsequently adopted in the present study.

These delay risk sources relate to: (1) owner; (2) consultant; (3) contractor; (4) design; (5) labor; (6) material; (7) equipment; (8) project; and (9) external aspects, and cover all possible sources of delay-inducing risk factors. Table 1 shows the different subsets of the 10 articles on which all the nine risk sources, and their constituting factors, were based.

Inevitably, variations existed in the individual delay risk factor lists identified by the reviewed articles within the different risk source categories. This was attributed to dissimilarities within the different articles in terms of construction environments, geographical conditions, political situations, construction methods, resource availabilities, and stakeholder engagements. As such, a *first series* of meetings were held with 15 construction experts to confirm the relevance of the identified risk factors to the construction sector and modify them as necessary. Based on the literature survey and the expert meetings, 59 delay risk factors were identified by the present study and a full listing of these factors within their source categories is presented in Table 2.

Data Set Compilation and Preprocessing

The data set used to develop the proposed predictive analytics tool included data from 51 construction projects, from 28 firms, that

Table 2. Complete list of identified risk factors within respective risk sources

Risk source	Identified risk factors
1. Owner	1.1 Inadequate project planning by owner
	1.2 Selecting inappropriate contractors
	1.3 Delays in site delivery to contractor
	1.4 Delays in reviewing and approving design documents
	1.5 Change orders by owner
	1.6 Slow decision-making process by owner
	1.7 Delays in progress payments by owner
	1.8 Suspension of work by owner
	1.9 Poor coordination by owner between consultant and contractor
	1.10 Conflicts between joint ownership of the project
2. Consultant	2.1 Delays in reviewing and approving design documents
	2.2 Delays in performing inspection and testing
	2.3 Delays in approving major changes in scope of work by consultant
	2.4 Inadequate consultant experience
	2.5 Poor consultant communication with contractor and owner
	2.6 Conflicts between consultant and design engineer

Table 2. (*Continued.*)

Risk source	Identified risk factors
3. Contractor	3.1 Ineffective project planning by contractor
	3.2 Difficulties in financing project by contractor
	3.3 Incompetence or inexperience of contractor
	3.4 Inadequate site investigation
	3.5 Slow site mobilization
	3.6 Poor site management and supervision
	3.7 Delays due to unreliable subcontractors' work
	3.8 Frequent change of subcontractors
	3.9 Rework due to errors during construction
	3.10 Poor contractor communication with consultant and owner
	3.11 Conflicts between contractor and consultant and/or owner
4. Design	4.1 Inadequate design team experience
	4.2 Misunderstanding of owner's requirements by design engineer
	4.3 Delays in producing design documents
	4.4 Design errors/incomplete or unclear design drawings
5. Labor	5.1 Shortage of labor
	5.2 Unqualified or inadequate workforce
	5.3 Low productivity of labor
	5.4 Personal conflicts among labor
6. Materials	6.1 Shortage of construction materials in market
	6.2 Delays in delivery of materials
	6.3 Inadequate quality of materials
	6.4 Damage of sorted materials
	6.5 Changes in material types and specifications during construction
7. Equipment	7.1 Shortage of equipment
	7.2 Slow mobilization of equipment
	7.3 Low productivity and efficiency of equipment
	7.4 Frequent equipment breakdowns
	7.5 Improper equipment or lack of high-tech equipment
8. Project	8.1 Unsuitable type of project bidding and award (e.g., negotiation, lowest bidder, etc.)
	8.2 Mistakes or discrepancies in contract documents
	8.3 Original contract duration is too short
	8.4 Ineffective delay penalties
	8.5 Lack of communication between project parties
	8.6 Legal disputes between project participants
9. External	9.1 Delays in obtaining permits from municipality
	9.2 Changes in government regulations and laws
	9.3 Delays in providing services from utilities (e.g., water, electricity, telephones, etc.)
	9.4 Unexpected surface and subsurface conditions (e.g., soil, water table, etc.)
	9.5 Problems with neighbors
	9.6 Unfavorable weather conditions
	9.7 Accidents during construction
	9.8 Price fluctuations

experienced varying degrees of time delay. Each data record in the data set represents a specific project and is linked to 10 data variables. The first of these variables represents the dependent variable that is the extent of TO sustained by the project. The other nine variables, reflecting the nine different delay risk sources aforementioned, are considered independent variables. Table 3 shows a sample of the described data set after all the operations pertaining to data collection and preprocessing (discussed next) were performed.

Table 3. Compiled data set structure

Project ID	Extent of TO	Risk source 1: owner	Risk source 2: consultant	…	Risk source 9: external
Project 1	30%–60% TO	Moderate	Very low	…	High
Project 2	>60% TO	Moderate	Very low	…	Very high
Project 3	>60% TO	Very high	Very low	…	High
⋮	⋮	⋮	⋮	⋱	⋮
Project 50	30%–60% TO	Very high	Very low	…	Very low
Project 51	>60% TO	High	Very low	…	High

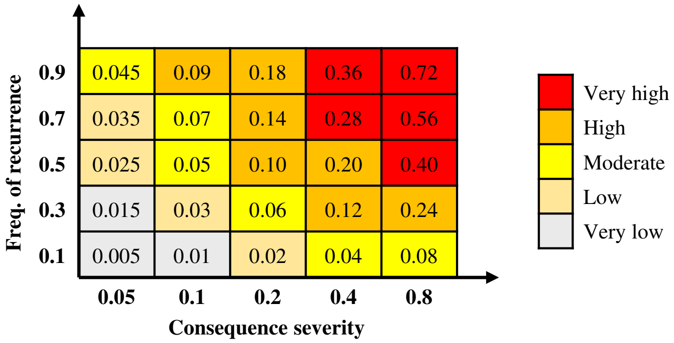
Table 4. Consequence severity index scale

Index value	Description
0.05	Contributes to no or insignificant time overrun
0.1	Contributes to <5% time overrun
0.2	Contributes to 5%–10% time overrun
0.4	Contributes to 10%–20% time overrun
0.8	Contributes to >20% time overrun

Table 5. Frequency of recurrence (throughout the project lifecycle) index scale

Index value	Description
0.1	Nonexisting or very rare
0.3	Rare
0.5	Moderate
0.7	Frequent
0.9	Very frequent

To compile the described data set, a *second series* of meetings were arranged. In total, 112 meetings were held across the 28 contacted firms over the span of 9 months to extract data from the 51 completed projects. The preliminary meeting with each firm involved explanations of the research goal and significance and the type of project data sought. Subsequently, one to two follow-up meetings were allocated to each project for collecting the necessary data. In these meetings, various project-related documents from the firm's historical records were investigated, including contract documents, specifications, change orders, schedule baselines, monthly and quarterly updates, resource calendars, and risk registers. Based on these records and the knowledge of the risk factors constituting the risk sources, each risk source was assigned scores (index values) on two different index scales. The first of these scales relates to the consequence severity toward affecting the project time objective, and the second relates to the frequency of recurrence throughout the project, as shown in Tables 4 and 5, respectively. The overall risk source contribution values toward the TO is then numerically evaluated through multiplying the two corresponding scores (Assaf and Al-Hejji 2006; Ismail et al. 2014; Xia et al. 2017). Afterward, these numerical risk source contribution values were discretized into categorical risk source contribution levels based on the discretization matrix shown in Fig. 2, where the values of the consequence severity are set along the horizontal axis, and those of the frequency of recurrence are set along the vertical axis. In this manner, each independent variable was classified into one of the following five categorical levels in terms of its contribution to TO: (1) very low; (2) low; (3) moderate; (4) high; and (5) very high. The two index scales and the discretization matrix were first adopted from literature (Assaf and Al-Hejji 2006; Mahamid 2011; Ismail et al. 2014; Kerzner 2017; PMI 2017; Xia et al. 2017), and then adapted based on input from the experts of the first series of meetings, and ultimately confirmed for adequacy to the projects investigated during the second series of meetings.



Freq. of recurrence					
0.9	0.045	0.09	0.18	0.36	0.72
0.7	0.035	0.07	0.14	0.28	0.56
0.5	0.025	0.05	0.10	0.20	0.40
0.3	0.015	0.03	0.06	0.12	0.24
0.1	0.005	0.01	0.02	0.04	0.08
	0.05	0.1	0.2	0.4	0.8
	Consequence severity				

Fig. 2. Discretization matrix for assessing risk source contribution level toward time overrun.

Furthermore, the dependent variable (i.e., TO) was also categorized into three class labels. One approach for such categorization is to divide the variable based on its frequency distribution into three *equal* class labels with each containing one third of the variable count (i.e., exactly 17 project records). This approach would yield three class labels, which are: (1) <21% TO; (2) 21%–27% TO; and (3) >27% TO. Although this categorization approach reduces bias when implementing predictive data analytics, it is atypical for practical applications. It would be more beneficial to categorize TO in a way that provides greater managerial insights and benefit. As such, the dependent variable was categorized into one of the following three class labels: (1) <30% TO; (2) 30%–60% TO; and (3) >60% TO, which resulted in the class labels containing 23, 16, and 12 project records, respectively. These three classes describe the extents of TO as minor, moderate, and major, respectively, and were agreed on during the second series of meetings with construction firms.

Data Set Constraints and Analysis Considerations

It should be noted that several constraints, considerations, and limiting factors were enforced and/or encountered during the data collection. First, the data gathered pertain to construction projects limited to those within Egypt and are owned, financed, designed, built, managed, and operated by national firms and entities. Second, the data collected were constrained to only building projects and do not concern other construction project categories (e.g., bridges) or disciplines. The third constraint pertains to project size, where data were only assembled from projects with contract values between 40 and 80 million Egyptian pounds, and with contract durations within the range of 1–2 years. Fourth, data were only collected from projects that had incurred time overruns (i.e., TO > 0). Unifying the type of construction data within the compiled data set was a key consideration for ensuring input consistency and homogeneity to the eventual predictive data analytics tool for attaining more reliable results, as mentioned earlier.

Furthermore, to ensure data integrity and quality, several carefully contemplated considerations were adopted. In this respect, only contractors with valid registrations in the Egyptian Federation for Construction and Building Contractors were considered. In addition, special attention was paid to consider only project cases having sufficient records and evidence to accurately infer credible and reliable data. Another important aspect was to avoid collecting data from projects with time intervals spanning across the period of the Egyptian revolution, which broke out on January 25, 2011. Such projects would have experienced relatively extreme conditions and would disturb the consistency of risk levels among the wider apprehension of assembled projects.

These constraints and considerations resulted in including only 51 project data records for further analyses, which, although constituted a valuable data set of variables that is rarely reported or studied in literature, was nonetheless relatively small. As such, the selection of the predictive data analytics tool used in the present study was based on tools that performed well on small-sized data sets, as will be explained in the "Selection of Predictive Analytics Tool and Algorithms" section.

Exploratory and Sensitivity Data Analyses

In order to better understand the system of delay-inducing risks within the construction sector, an exploratory data analysis, the results of which are displayed in Fig. 3, was conducted to visually represent the properties of the collected data records and the correlations between different variables. The figure is a 10 × 10 matrix, in which the class-labeled dependent variable (i.e., TO) and the nine independent variables (i.e., risk sources) are shown on the rows and columns. To facilitate its interpretation, the figure is divided into five blocks: the leftmost column—Block 1; the diagonal—Block 2; the uppermost row—Block 3; the lower-left triangle—Block 4; and the upper-right triangle—Block 5. The exploratory data analysis provides *three main insights*: (1) variable frequency counts and smoothed frequency curves; (2) a sensitivity analysis that explores the dependency of TO on the different risk sources; and (3) a sensitivity analysis that explores the dependencies of the different risk sources on one another.

Regarding the *first insight*, the boxes in Blocks 1 and 2 show the frequency counts and the smoothed frequency curves, respectively, of the data variables within each class. As mentioned earlier, the counts of projects incurring <30% TO, 30%–60% TO, and >60% TO were 23 (45%), 16 (31%), and 12 (24%), respectively. These percentages are consistent with the prevailing phenomenon of building projects experiencing less serious delays more frequently than more serious delays (Abd El-Razek et al. 2008; Singh 2009). It can also be seen that the risk contribution values of the owner, contractor, project, and external risk sources are higher than those related to other risk sources.

The *second insight* is inferred through the box plots located within Block 3 that demonstrate how the changes of TO from one class to another are sensitive to/affected by any changes in the risk contribution values of different risk sources. These box plots illustrate the risk contribution value ranges and distributions pertaining to the nine risk sources for each class of TO, where the thick black bars and the middle boxes represent the median values and the interquartile ranges of these distributions, respectively. An important observation is that, for each of the nine risk sources, the risk value distributions for each class are highly overlapping. This overlapping indicates that no individual risk source is fit to provide a definite/clear distinction of the TO class, which subsequently entails that the risk sources are heavily interdependent with regard to influencing TO, and that the studied system exhibits a significant level of complexity. Nonetheless, some findings from the box plots will be described for completeness. Most notably, higher project risk values are associated with the <30% TO class, while lower risk values are not distinctly related to a certain TO class. This implies that minor TO extents are sensitive to project-related risks, whereas moderate and major TO extents are insensitive. Moreover, lower owner risk values are associated with the <30% TO class, whereas with higher risk values, 30%–60%, and >60% TO extents tend to occur. Accordingly, for the owner-related risks, minor, moderate,

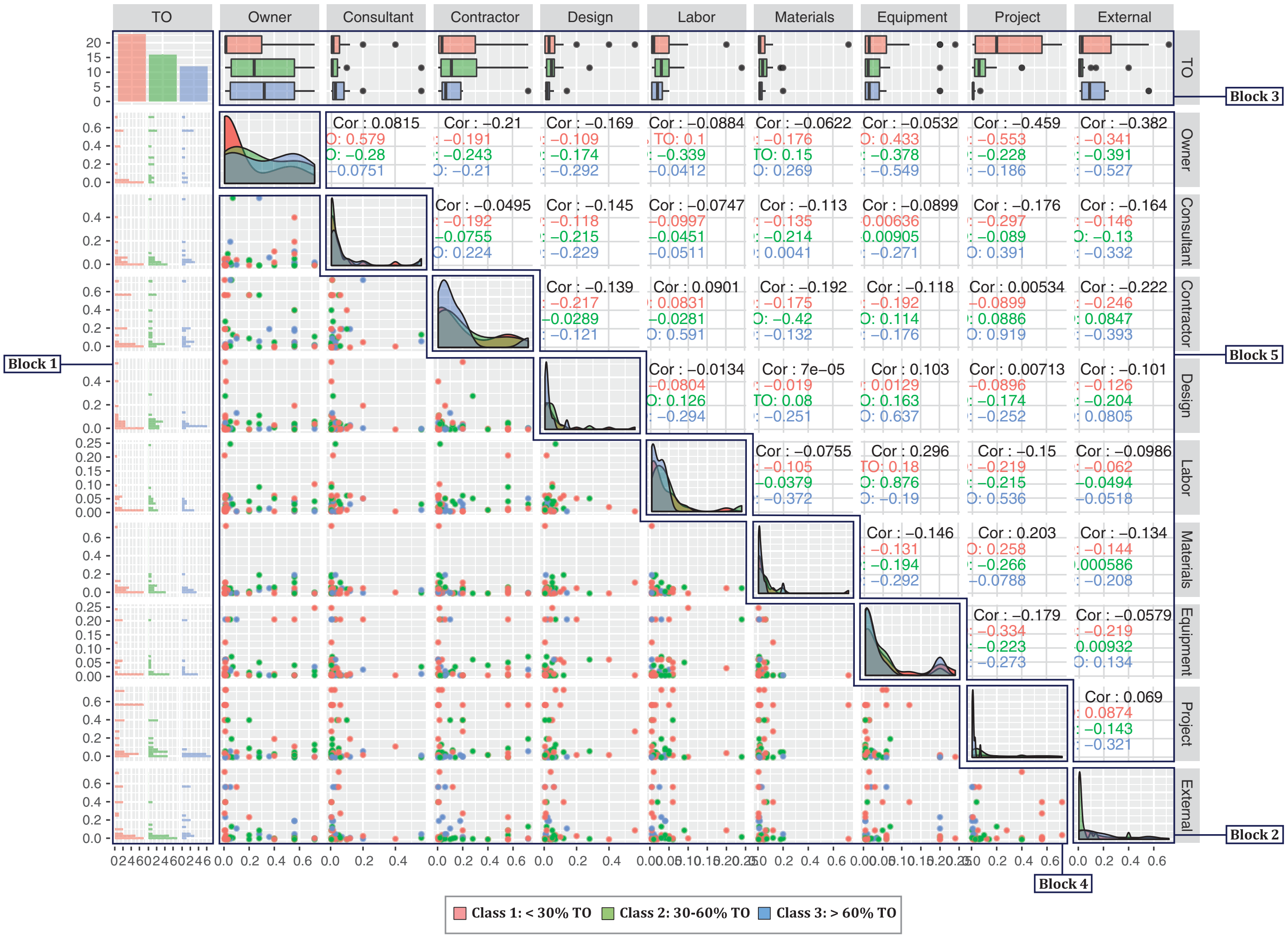


Fig. 3. Exploratory and sensitivity data analysis results of time overrun extent and the nine risk sources.

and major TO extents are all sensitive; however, there is no clear separability/discrimination between the latter two since the risk value distributions for these two classes are nearly identical and almost fully overlapping. Through similar interpretations, it can be noted that for contractor risks, minor and moderate TO extents are sensitive albeit with a lack of clear separability between the two classes, and that major TO extents are insensitive. In addition, for external risks, only minor TO is sensitive. A final observation is that TO is insensitive to consultant, design, labor, and materials risks, which is also attributed to the fact that these risk ranges are all limited to lower risk values.

Regarding the *third insight*, Blocks 4 and 5 represent a correlation matrix that illustrates the dependencies among the risk sources. The boxes in Block 4 show scatter plots of the risk values of every two risk sources. Complementing these scatter plots, the boxes in Block 5 show the corresponding correlation values (ranging from +1 to −1), which describe the strength and direction of the relationship between the risk sources, both for all project records collectively and for project records within each TO class separately. The magnitude of the correlation value is indicative of the strength of the relationship between any two risk sources (i.e., how much variance in one data variable is explained by the variance in the other data variable), whereas the sign of the correlation value specifies the direction of this relationship (i.e., whether the data variable values vary together positively or negatively). As expected, no strong positive or negative correlation exists among any pair of risk sources, either for all project records collectively or for project records within a specific TO class separately. This implies that: (1) for the considered data set, no two risk sources are related to one another, adding to the complexity of the system; and (2) the property of variable conditional independence for each class is manifested in the current data set of project records. The latter finding played an important role in selecting one of the predictive analysis approaches, as will be discussed in the following section.

The *three key conclusions* from the exploratory and sensitivity data analyses conducted in this section can be summarized as: (1) delay risk sources are highly interdependent, which is evident from the lack of any specific dependence trend of the TO on any individual risk source; (2) delay risk sources and TO classes are related in a complex manner, which is evident from the class inseparability issue; and (3) delay risk sources are also, among themselves, related in a complex manner, which is evident from the weak correlations between pairs of risk sources. These conclusions demonstrate the complexity of the studied system (i.e., the manner in which the independent and dependent variables are altogether related), thereby asserting the complex nature of the construction sector and its inherent delay-inducing risks. It is this complexity that guided the selection of a capable predictive data analytics tool to realize an accurate predictive delay risk analysis model, as will be discussed in the next section.

Selection of Predictive Analytics Tool and Algorithms

Tool Selection

ML is one of the most promising tools in predictive data analytics. It combines methods from statistics, database analysis, data mining, pattern recognition, and artificial intelligence to extract trends, interrelationships, patterns of interest, and useful insights from complex data sets (Aburrous et al. 2010; Flath et al. 2012). In the present study, ML was selected over other predictive data analytics tools, such as, for example, statistical learning (SL), for two main reasons.

First, as detailed by Breiman (2001), Tixier et al. (2017), and Dindarloo and Siami-Irdemoosa (2016), SL-based models require both formal model structures and data frequency distributions to be imposed a priori to the data fitting processes either based on some knowledge of the system or arbitrarily through assumptions. However, data generated from complex systems (such as those analyzed herein) would rarely have model structures and frequency distributions that are known or tractable. To reiterate, based on the conclusions of the exploratory and sensitivity data analyses conducted in the previous section, the complex nature of the studied system was apparent through: (1) the highly interdependent delay risk sources, which lacked any specific trend for the subsequent TO extent—indicating that the underlying model structure is complex for such a system; and (2) the complex relationships between the risk sources, both with the TO classes and among themselves—indicating that the underlying variable frequency distributions for such a system are intractable.

The power of ML algorithms relies on their ability to avoid the limitations of any explicitly programmed instructions concerning the model structure and any hypothetical assumptions pertaining to the data frequency distributions. In fact, the underlying ML assumption is that the forms by which the independent variables and dependent variable are altogether related are complex and unknown. ML thus focuses on learning from the implicit data patterns through algorithms that continuously improve their performances through experience and induction. Based on the aforementioned, ML algorithms are not only effective in dealing with data variables having simple linear or nonlinear relationships, but also with variables having complex high-order relationships, or even disjunctive variables. It can thus be argued that the adoption of conditioned analysis methods (e.g., SL) to analyze data collected from complex systems may result in imposed model structures and/or data frequency distributions that are poor representations of the actual system's phenomena. Such adoption would undermine the predictive accuracy of the resulting model compared to the models to be generated by the more versatile ML approach.

Second, ML uses optimization techniques to maximize the predictive performances (by minimizing the number of incorrect predictions) of the generated models, while SL focuses on the inferences induced from the relationships between the variables in the statistical model. Therefore, SL-based models typically face deficiencies in their predictive performances when dealing with data sets having a large number of variables, while ML-based models are known to be more suited to analyze such data sets and typically yield higher predictive accuracies (Kim et al. 2008; Aggarwal 2016). As the data collected within this study consist of nine independent variables (risk sources) and one dependent variable (TO), and based on the complexity of the system which was discussed in the "Exploratory and Sensitivity Data Analyses" section, ML was adopted in the present study.

ML Algorithms Selection

The R open source platform (R Core Team 2013) was used in the present study and is a powerful computation tool that supports implementations of different ML techniques such as clustering, classification, and regression. Classification is a supervised ML technique that is very effective in predictive data analytics. It is based on learning, via historical/training data, to facilitate mapping new input records (e.g., project cases) into specific dependent variable output classes (e.g., the extent of TO) based on relevant independent variable values (e.g., project-anticipated risk contribution levels). The present study focused on using the classification technique because of its capability of handling complex-related variables and its effectiveness in dealing with categorical variables (Aggarwal 2016).

Two classification algorithms were applied in this study, which are the decision tree (DT) and naïve Bayesian (NB) algorithms. Through a review of the different ML classification algorithms reported in the literature, these two algorithms were selected mainly, among other reasons, because they are suited to small-sized data sets with a demonstrated history of satisfactory performance (Amor et al. 2004; Chi et al. 2012; Ashari et al. 2013; Aggarwal 2016).

The DT classification algorithm produces an indicative classifier/model for segmenting new data records into class labels by modeling the classification process through a set of hierarchical decisions (rules) concerning the data variables. The induced decisions are arranged in a tree-like structure that is initiated by identifying the root node and then recursively splitting nodes until no further divisions are possible. The splitting criteria are derived from concepts of information theory, which depend on the values of information gain, or entropy reduction, to assess the amount of information needed to generate decisions for segmenting a data record into a class label. Such information theoretic measures are key as they represent the criterion for assessing the hierarchical order of variables along the tree and the splitting of the nodes throughout (Caldas and Soibelman 2003; Desai and Joshi 2010; Aggarwal 2016). Different available ML algorithms in the R platform can be used to develop DT classifiers. Some examples of such algorithms include ID3 (Glur 2018), rpart (Therneau and Atkinson 2018), tree (Ripley 2018), J4.8 (Hornik et al. 2009), and C5.0 (Kuhn and Quinlan 2018). The recursive partitioning and regression trees (rpart) algorithm in R was selected within the present study. The rationale behind selecting this algorithm is its proven robustness against noisy data, its capability of learning disjunctive variable relationships, and its proven performance with small data sets (Chi et al. 2012; Aggarwal 2016).

The NB classification algorithm, on the other hand, produces a classifier that identifies classes for new data records by calculating joint conditional probabilities of the previous data records' independent variable values given their dependent variable class labels. This algorithm is based on Bayes' theorem which quantifies the conditional probability of random variables (Gong et al. 2011). It also assumes that the naïve assumption, that variable values are conditionally independent for each class, holds (Ng and Jordan 2002; Aggarwal 2016). The outputs of the produced model are conditional probability scores and mutually exclusive class label designations based on the highest class label joint probability value for the data record (Gong et al. 2011; Bilal et al. 2016; Aggarwal 2016). The naïve Bayes algorithm in R (Meyer et al. 2017) was selected because it is ideal for small-sized data sets since it is known to converge quicker than other algorithms and, as such, requires less training data (Amor et al. 2004; Ashari et al. 2013). It should be noted, however, that NB algorithms are only suitable for analyzing data sets with conditionally independent variables, which was evident from the properties of the data set considered in the present study, as explained previously in the "Exploratory and Sensitivity Data Analyses" section.

In terms of DT or NB previous applications within the construction research field, Caldas and Soibelman (2003) presented a model based on automatic hierarchical classification to enhance the access and organization of unstructured text documents within construction management information systems. Another study, carried out by Desai and Joshi (2010), utilized a DT classification mining algorithm to assess the most important factors influencing labor productivity in Indian construction projects. In addition, a study by Chi et al. (2012) applied four different DT classification algorithms to predict the cost performance of projects. Furthermore, an application of classification and regression trees was presented by Chou and Lin (2013) to proactively forecast disputes in the initiation phase of public–private partnership projects.

Jiang and Mahadevan (2008) proposed a Bayesian probabilistic methodology to assess the nonparametric damage detection of building structures. Bayesian learning methods were also used by Gong et al. (2011) for identifying and classifying worker and heavy equipment actions in challenging construction environments from video data sets. Moreover, Bayesian networks were applied to analyze the specific causes of different types of accidents associated with the construction of embankments by Gerassis et al. (2017). Evidently, and to the best of the authors' knowledge, it can be seen that, although ML classification algorithms are widely used in various disciplines, their applicability has rarely been exploited in the construction sector, particularly in the delay risk analysis area.

Decision Tree and Naïve Bayesian Classifiers Training

This section focuses on describing the analyses performed to achieve the second objective of the present study (Fig. 1). Both DT and NB classification algorithms were used to generate classifiers that predict the time performance of projects based on their risk source levels by partitioning each project into a class label describing the expected time delay. Each algorithm initially defines the data set as an information system with a finite set of data records and variable values. Each row is considered a distinct project record, and each column is considered a distinct variable of that record. The model then identifies the independent variables (the nine risk sources) and the single dependent variable (TO). When implementing supervised classification learning, the ML algorithms are used to learn the internal structure of the data set to examine the effect of the variations of different variables on the degree of TO sustained. Each algorithm then forms a classifier to predict class labels for any new records.

Decision Tree Classifier

As previously discussed, the DT algorithm analyzes the training data for learning the influences of independent variables on partitioning data records into class labels. The algorithm then outputs a classifier in the form of a tree-like structure that describes the decision flow.

First, information theoretic measures of entropy and gain for all independent variables are calculated to be used as criteria for tree construction and node splitting. The classifier considers an independent variable to be more informative, and thus affects the dependent variable more significantly, when more information is induced by knowing the variable's value for predicting the dependent variable's class label. In other words, an independent variable is relevant if by eliminating knowledge about this variable, estimating the dependent variable can be substantially adversely affected. The algorithm then builds on these information theoretic insights and develops a knowledge-inductive decision tree for partitioning new records into predefined classes through conjunctive if-then rules.

Applying the DT algorithm to the described data set used in this study generates the decision tree structure shown in Fig. 4. The decision tree grows from the top node, referred to as the root node, and forms a hierarchical structure to map new data records (representing risk source levels of new construction projects) into class labels (describing the project's expected extent of TO). Apart from the root node, the tree consists of internal nodes, leaf nodes, and branches. In general, nodes represent class labels, and branches refer to the associated variables and variable values. Starting from the root node, data are recursively split to form new tree levels, where each level comprises internal nodes connected by branches. This splitting criteria is based on the previously explained information theoretic measures, where each possible split of the data is examined at each node, and the variable with the highest information gain (i.e., highest influence on TO) is chosen for splitting the data. Accordingly, the branches represent possible variable values from the node that they originate.

In that sense, each node represents a subset of the data space defined by the combination of split criteria in the nodes above it, which is why the root node is the only node corresponding to the entire feature space. In addition, each node is labeled with the dominant class according to the distribution of classes within the node. Nodes also return information regarding this class distribution in the form of counts and percentages of the class labels at that node. This recursive partitioning process continues until there is no more benefit from further data segmentation, signaling that additional tree levels would cause data overfitting that would lead to higher levels of misclassification, thus diminishing the algorithm's predictive performance. Nodes that are at the end of the last branches on the tree are called leaf nodes and play an important role when the tree is used as a predictive model. These leaf nodes represent the outcomes of all prior decisions and refer to the class label that all data records following the path to that leaf would be segmented.

By referring once again to the generated decision tree from the 51 compiled projects data set in Fig. 4, the following observations can be made. Each node refers to one of the three classes: Class 1 (<30% TO); Class 2 (30%–60% TO); or Class 3 (>60% TO). Each branch represents, out of the nine independent variables, the selected risk source to be split, and indicates its variable levels (very low, low, moderate, high, or very high). The instances of project TO at the root node are all the instances in the data set. As such, this root node contains 51 instances, of which 23 projects experienced TO of less than 30%, 16 projects between 30% and 60%, and 12 projects greater than 60%, as explained earlier. Consequently, the root node is labeled as a "Class 1" node. Furthermore, the hierarchical order of variables along the tree and the splitting of each node follow the criteria of information entropy and gain as previously discussed. The top three levels of the tree are occupied by the project, owner, and contractor sources, respectively, indicating their high significance toward influencing TO. Similar interpretations can be deduced from the remaining nodes and branches of the decision tree.

For further interpretation of the decision tree logic, the tree can be converted into a set of rules to be used for predicting the time performance of new project cases. Rules are generated by traversing each branch of the tree and collecting the variable values until a root node is encountered, indicating the predicted class label. A confidence percentage is associated with each generated rule and

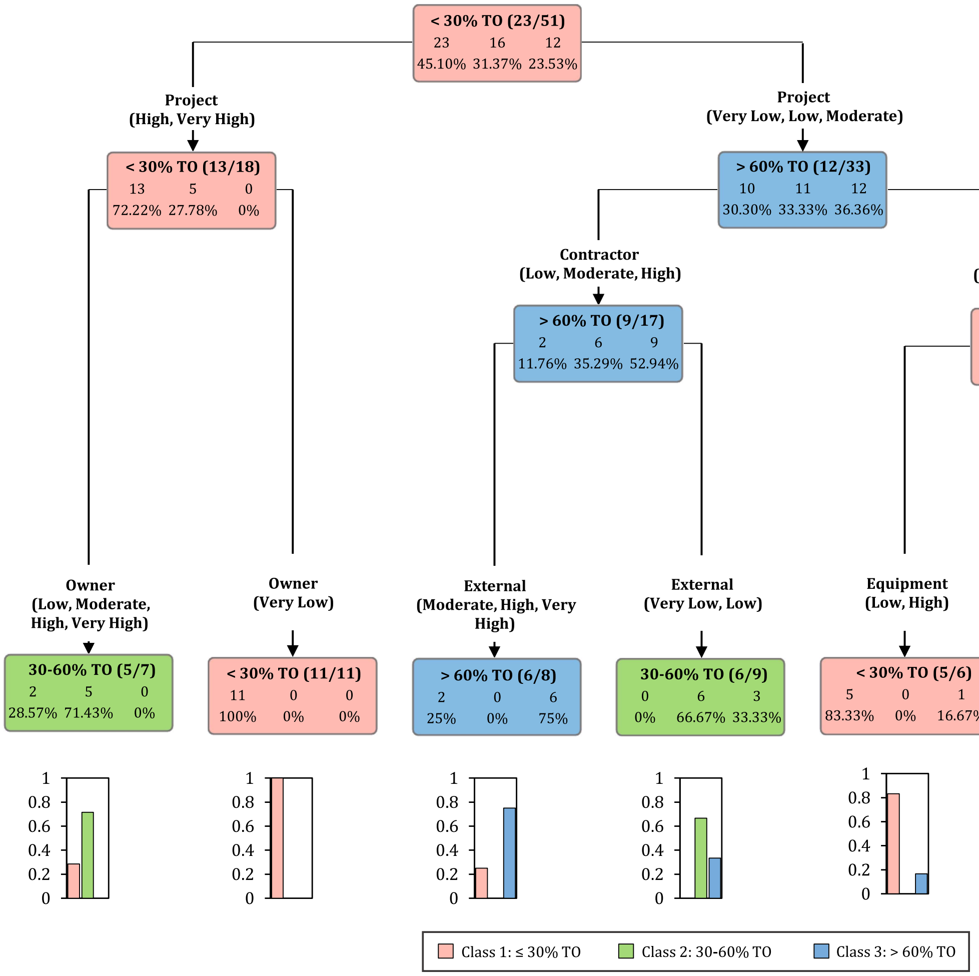


Fig. 4. Decision tree for predicting project time overrun from risk source levels.

describes the confidence in the class predicted by the rule (shown as a probability score in the leaf node). The model produced seven rules that follow a series of logical *if-then* statements. The rules produced from Fig. 4, taken from top to bottom and from left to right, are shown in Fig. 5. Interesting predictive patterns can be derived from these rules, and for further clarification, Rules 1 and 2 will be discoursed. Rule 1 states that for a project with high or very high levels of project-related risk factors, as well as low, moderate, high, or very high levels of owner-related risk factors, there is a 71.4% chance that the project will be delayed beyond completion date by 30%–60% of its original project duration. Rule 2 indicates that for a project with high or very high levels of project-related risk factors, and a very low level of owner-related risk factors, it is certain that the project will experience a TO of less than 30%. In a similar manner, interpretations can be deduced from the remaining rules.

Naïve Bayesian Classifier

The NB classifier identifies mutually exclusive classes of TO for new project cases through calculations of conditional probabilities of variable values with relation to their class labels. It is based on the Bayes theorem, which quantifies the conditional probability of a random variable, and on the naïve assumption of variable conditional independence. To describe the NB algorithm, the Bayes law must first be introduced for completeness. The Bayes law is shown in Eq. (1), where C is the class label, A is the variable value of the new data record, and $P(C|A)$ is the conditional probability of C given that A is observed. The Bayes theorem is useful for estimating $P(C|A)$ when it is difficult to be attained from the training data, but other values as $P(A|C)$, $P(C)$, and $P(A)$ can be obtained more easily

$$P(C|A) = \frac{P(A|C)P(C)}{P(A)} \tag{1}$$

A more realistic approach would be to consider the case in which a data record has several (m) independent variable values, $A = (a_1, a_2, \ldots, a_m)$. The objective is to assign this record to a definite class C_i (which is one of n class labels) such that it corresponds to the maximum value of $P(C_i|A)$. In that sense, Eq. (2) could be inferred. It is based on the product of conditional probabilities of independent variable values $a_1, a_2, \ldots, a_m$ given that class C_i is observed

$$P(C_i|A) = \frac{P(a_1, a_2, \ldots, a_m|C_i)P(C_i)}{P(a_1, a_2, \ldots, a_m)} = \frac{(\prod_{j=1}^{m} P(a_j|C_i)) \times P(C_i)}{P(a_1, a_2, \ldots, a_m)} \quad \text{where } i = 1, 2, \ldots, n \tag{2}$$

Since the denominator is independent of the class, it thus suffices to only compute the numerator value in order to determine the class with maximum $P(C_i|A)$. Therefore, the NB model equation is

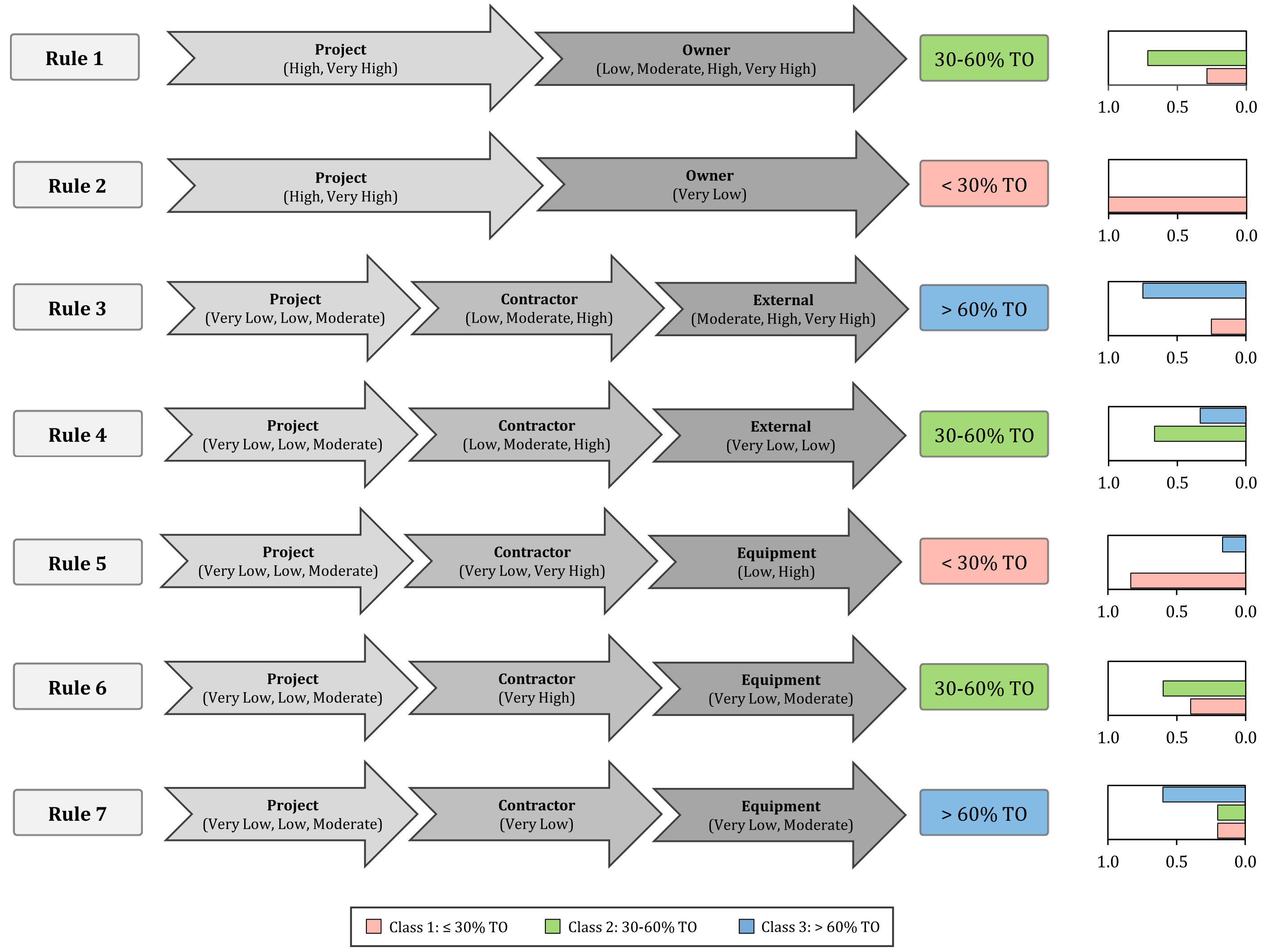

Rule 1
Project (High, Very High)
Owner (Low, Moderate, High, Very High)
30-60% TO
Rule 2
Project (High, Very High)
Owner (Very Low)
< 30% TO
Rule 3
Project (Very Low, Low, Moderate)
Contractor (Low, Moderate, High)
External (Moderate, High, Very High)
> 60% TO
Rule 4
Project (Very Low, Low, Moderate)
Contractor (Low, Moderate, High)
External (Very Low, Low)
30-60% TO
Rule 5
Project (Very Low, Low, Moderate)
Contractor (Very Low, Very High)
Equipment (Low, High)
< 30% TO
Rule 6
Project (Very Low, Low, Moderate)
Contractor (Very High)
Equipment (Very Low, Moderate)
30-60% TO
Rule 7
Project (Very Low, Low, Moderate)
Contractor (Very Low)
Equipment (Very Low, Moderate)
> 60% TO
1.0 0.5 0.0
🟥 Class 1: ≤ 30% TO
🟩 Class 2: 30-60% TO
🟦 Class 3: > 60% TO

Fig. 5. Decision rules for predicting project time overrun from risk source levels.

simplified by removing the denominator as it will have no impact on the conditional probability outcome:

$$P(C_i|a_1, a_2, \ldots, a_m) \propto \left(\prod_{j=1}^{m} P(a_j|C_i)\right) P(C_i)$$

$$\text{where } i = 1, 2, \ldots, n \tag{3}$$

It can be interpreted from Eq. (3) that a data record with variable values $A = (a_1, a_2, \ldots, a_m)$ is allocated to a class label C_i, which returns the highest value of $P(C_i|a_1, a_2, \ldots, a_m)$, which is proportional to the product of the various $P(a_j|C_i)$ multiplied by the probability of that class label existing in the data set, which is $P(C_i)$.

Upon application to the data set, the model outputs are conditional probability scores of each independent variable level for each risk source given each of the three dependent variable class labels. The results of the first risk source (owner risk source) are shown in Table 6 as a sample. For any new project case, the model computes values of conditional probability products for its independent variable levels given each class label, multiplied by the probability of retrieving that class from the data set. Subsequently, the model maps this project to its predicted class label based on the maximum of the three values corresponding to each of the classes. Evaluations of the NB classifier's predictive accuracy and performance comparisons with the DT classifier will be discussed next.

Table 6. Class conditional probabilities for owner risk source

	Variable level				
Class label	Very low	Low	Moderate	High	Very high
<30% TO	0.609	0.043	0.000	0.087	0.261
30%–60% TO	0.125	0.063	0.188	0.125	0.500
>60% TO	0.167	0.000	0.167	0.083	0.583

Model Performance Evaluation and Validation

After introducing and applying both the DT and NB classification models, the purpose of this section is to validate their predictive performance and evaluate/compare the effectiveness of both models in analyzing the project data set. Primarily, references from the literature will be called upon to develop a well-rounded interpretation of the common performance of both models based on different domain aspects. Generally, NB classifiers do not require a large amount of data to acquire the internal structure of a data set, and are therefore better than DT classifiers in learning from smaller training sets while reaching high levels of classification accuracy (Ashari et al. 2013). Moreover, from a computational perspective, NB classifiers are typically faster and more efficient in terms of both their learning and predictive capabilities (Amor et al. 2004). However, NB classification follows the laws of independent events' probability; hence, a central assumption in applying NB classifiers is that for each class, variable values are all conditionally independent of one another. Therefore, DT classifiers typically perform better in domains involving correlated variables. In other words, if two or more variables are highly correlated in NB classification, more weight is allocated to their influence on the predicted class label, which leads to a decline in predictive accuracy. DT models do not suffer from such an undesirable bias because it would not be possible to use two correlated variables for splitting the data of the training set, since this would lead to exactly the same split (Xhemali et al. 2009; Niuniu and Yuxun 2010). It has been discussed that the considered project data set is relatively small in size and the projects have variables that are conditionally independent of one another. For these two reasons, and based on the overall inferences presented previously, the authors' preliminary hypothesis was that the NB model would generally outperform the DT model for a data set of such properties.

Performance Evaluation Indices

The models' performance evaluations are facilitated by further developing the algorithms in R to return confusion matrices. Confusion matrices are specific table representations that describe the performance of classification models. The confusion matrices of the DT and NB classifiers are shown in Tables 7 and 8, respectively, where the classifiers were initially deployed to train on the entire data set. Confusion matrices include integers reflecting the counts of certain classifications. Rows correspond to the number of actual classifications or total number of records within each class, while columns represent the number of predicted classifications. All correct predictions are located in the diagonal of the table, and this facilitates the visual inspection of the matrix for errors, which are any nonzero values outside the diagonal. Before proceeding to evaluate the model performances from the matrices, a few key terms for each class need to be clarified first:

1. True positives (TPs): Number of predictions that were correctly assigned to a class (i.e., value in the matrix diagonal for the corresponding class).
2. False positives (FPs): Number of predictions that were incorrectly assigned to a class (i.e., sum of values in the corresponding class column excluding the TPs).
3. False negatives (FNs): Number of predictions that were incorrectly unrecognized as class assignments (i.e., sum of values in the corresponding class row excluding the TPs).
4. True negatives (TNs): Number of predictions that were correctly recognized as not belonging to a class (i.e., sum of values of all rows and columns excluding the row and column of that class).

Table 7. Confusion matrix from DT classifications

	Predicted class			
Actual class	<30% TO	30%–60% TO	>60% TO	Totals
<30% TO	16	4	3	23
30%–60% TO	0	14	2	16
>60% TO	1	3	8	12
Totals	17	21	13	—

Table 8. Confusion matrix from NB classifications

	Predicted class			
Actual class	<30% TO	30%–60% TO	>60% TO	Totals
<30% TO	18	3	2	23
30%–60% TO	3	12	1	16
>60% TO	1	1	10	12
Totals	22	16	13	—

Based on the aforementioned terminologies, confusion matrices enable analysts to extract numerical measures that act as indicators of the model performance. Such measures could be either overall performance indices or class performance indices, since this is a multiclass classification.

The two model overall performance indices used in this study are *accuracy* and *misclassification error*. Accuracy is a percentage of total number of correct classifications to total number of predicted classifications by a model, and correspondingly, the misclassification error (also referred to as the error rate) is a percentage of the direct misclassifications. In other words, the overall accuracy can be perceived as the ratio between the sum of diagonal values and the sum of the table. Thus, the confusion matrix of a high-performing model has large numbers in its diagonal and small numbers (ideally zero) outside the diagonal. Model class performance indices include *precision*, *sensitivity*, *specificity*, *false positive rate* (*FPR*), and *false negative rate* (*FNR*), and are calculated from the confusion matrices as shown in Eqs. (4)–(8) respectively

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{6}$$

$$\text{FPR} = \frac{FP}{FP + TN} \tag{7}$$

$$\text{FNR} = \frac{FN}{TP + FN} \tag{8}$$

Validation Approaches

The validation of both models was performed based on two approaches that reflect the two stages of a typical ML procedure: training and testing. In the first approach, and due to the small size of the compiled project data set, training performance was evaluated by deploying the models to learn from the entire data set and, subsequently, predicted the class labels of the same data used for training. In the second approach, to evaluate the testing performance, the models were trained to learn from a subset of the entire data set (a training set) and then predict class labels for the remaining part of the data set (a testing set) in order to ensure an unbiased evaluation.

The holdout method is a common practice for investigating model testing performance, where the complete data set is randomly split into 80%–60% and 20%–40% portions for training and testing sets, respectively. However, the major drawbacks of the holdout method include difficulties with arriving at a random testing set split that would be representative of the entire data set in terms of: (1) the true variability of the independent variables; and (2) the distributions of the three class labels of the dependent variable in a way that avoids class imbalance (Kim et al. 2008; Chou and Lin 2013).

In this respect, k-fold cross validation was adopted for evaluating the testing performance since it is known to be a reliable method that minimizes the bias and variance associated with the random splitting performed in the holdout method (Kohavi 1995; Hastie et al. 2009; Arlot and Celisse 2010; Seong et al. 2018). In k-fold cross validation, the complete data set is divided into k distinct and almost equal subsets or folds, where k is a positive integer. The holdout method is then repeated k times; each time one of the k folds is held out rotationally as the test set and the other k-1 folds are put together for training. For every repetition, a confusion matrix is obtained from which overall and class performance indices can be extracted. The final k-fold cross validation performance indices are then computed by averaging these k individual indices (Gong et al. 2011; Son et al. 2014; Tixier et al. 2017). As such, the advantage of this method is that the entire data set is used in both training and testing whereas each fold, and hence each data record, is retained exactly once for testing. The present study used 10-fold cross validation because many researches have reported $k = 10$ to be optimal in terms of computational time, estimation of error, and variance of indices (Kohavi 1995; Hastie et al. 2009; Wei et al. 2013).

Analysis Results and Discussion

The training and testing performances of both DT and NB classifiers were evaluated based on the performance indices calculated from the confusion matrices. In general, unrepresentable model ability to predict the testing set is apparent from the results generated by the 10-fold cross validation. This is attributed to the small size of the data set (51 data records) where a single test fold contains 5–6 records and training folds sum up to 45–46 records. Small-sized testing and training sets often lead ML models to: (1) overfitting, where the models memorize the peculiarities of the training data rather than its general structure; and (2) generating performance indices with high variances among the test folds (Kim et al. 2008). Therefore, for small-sized data sets, it is important to note that testing performance indices may not reflect individual model performance but may rather serve the comparative study between the DT and NB models.

Table 9. Comparison between classifiers based on overall performance indices

Performance index	DT classifier	NB classifier
Accuracy	74.5% (47.2%)	**78.4% (51.2%)**
Misclassification error	25.5% (52.8%)	**21.6% (48.8%)**

Note: Values without parentheses are training performance indices; values in parentheses are testing performance indices; and values in bold indicate the superior performance.

Table 9 shows a comparison of the overall performance indices where the training (testing) values of accuracy and misclassification error for the DT classifier are 74.5% (47.2%) and 25.5% (52.8%), respectively, and for the NB classifier are 78.4% (51.2%) and 21.6% (48.8%), respectively. It can be inferred that both models perform reasonably well in terms of training, with the NB classifier exhibiting the relatively better performance in both training and testing abilities. Accordingly, the NB model is showing initial signs of exceeding the DT performance as suggested by the writers' hypothesis. Nevertheless, more performance measures need to be examined for a wider perspective.

Class performance indices are computed by both models for each of the three classes and Table 10 summarizes the comparison of training and testing abilities. Moreover, to enhance the visual interpretation of model performance by class, the class performance indices were plotted as shown in Fig. 6. Overall, the predictive performance of the NB classifier is largely impressive. Apart from its high training accuracy of 78.4%, its training measures of precision, sensitivity, and specificity do not fall below 75.0%, 75.0%, and 85.7%, respectively, and its measures of FPR and FNR do not exceed 14.3% and 25.0%, respectively. Furthermore, the results display the NB classifier's consistent performance throughout the classes. In addition, the results show that the DT model also performs reasonably well, in terms of training, with a relatively low misclassification error and rather good class performances. The DT classifier's least values of precision, sensitivity, and specificity are 61.5%, 66.7%, and 80.0%, respectively, and highest values of FPR and FNR are 20.0% and 33.3%, respectively. Similar findings and trends can be interpreted for the testing performances of both DT and NB classifiers.

In terms of comparing both models, it can be inferred that the NB classifier's predictive performance exceeds that of the DT classifier in terms of both training and testing capabilities. The described results indicate the superiority of the NB classifier with regard to the minimum threshold attained for precision, sensitivity, and specificity, and maximum threshold attained for FPR and FNR. Other important insights can be made by comparing the models' performance in each class. The NB classifier returns higher values in two out of three class comparisons concerning precision, sensitivity, and specificity. As for FPR and FNR comparisons, the NB model was also found to have a better performance since it returned lower values in two out of three class comparisons. It is also clear from the figures that the NB classifier displays a more consistent performance across the three classes compared to its DT counterpart. As such, the NB model outperforms the DT model in terms of overall performance, as well as in every class performance measure.

As a final statement, proactive project risk management entails the identification of new arising risk factors as well as the continuous monitoring of both the established and arising risk factors'

Table 10. Comparison between classifiers based on class performance indices

	DT classifier			NB classifier		
Performance index	<30% TO	30%–60% TO	>60% TO	<30% TO	30%–60% TO	>60% TO
Precision	**94.1%** (51.7%)	66.7% **(40.0%)**	61.5% (20.8%)	81.8% **(54.0%)**	**75.0%** (21.7%)	**76.9% (45.0%)**
Sensitivity	69.6% **(63.3%)**	**87.5%** (40.0%)	66.7% (35.0%)	**78.3%** (60.0%)	75.0% **(45.0%)**	**83.3% (45.0%)**
Specificity	**96.4%** (59.1%)	80.0% **(90.5%)**	87.2% (66.3%)	85.7% **(61.7%)**	**88.6%** (82.1%)	**92.3% (79.2%)**
False positive rate (FPR)	**3.6%** (40.8%)	20.0% **(9.5%)**	12.8% (33.7%)	14.3% **(38.3%)**	**11.4%** (20.3%)	**7.7% (20.0%)**
False negative rate (FNR)	30.4% **(36.7%)**	**12.5%** (60.0%)	33.3% (65.0%)	**21.7%** (40.0%)	25.0% **(55.0%)**	**16.7% (55.0%)**

Note: Values without parentheses are training performance indices; values in parentheses are testing performance indices; and values in bold indicate the superior performance per class label.

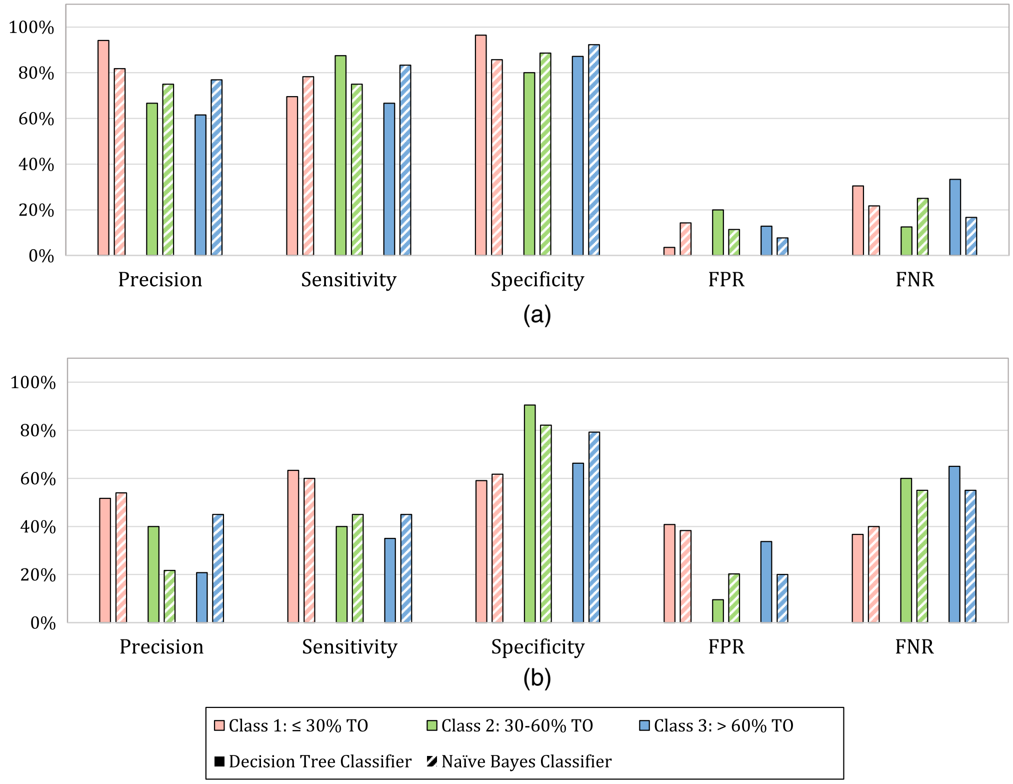


Fig. 6. Comparison between classifiers in terms of (a) training; and (b) testing performances based on class performance indices.

dynamic behavior throughout the project lifecycle. In this respect, such monitoring involves the timely tracking and reassessment of the risk sources' expected risk severity and recurrence scores and, hence, their overall risk contribution levels. As such, the long-term goal of the present research is to create an analysis platform that, through modifying the independent variable input risk values, would facilitate continuous refinement of project duration prediction throughout the project lifecycle. As a first and key step toward meeting this long-term goal, the focus of the present study was to create trained ML models that are capable of conducting such dynamic analysis when such dynamic data become available.

Conclusion

The construction sector is a knowledge-based domain that deals with large volumes of objective, heterogeneous, and interdependent data encapsulating abstract knowledge. In most cases, construction firms fail to capitalize on the opportunity presented by this data availability whereas, typically, conventional risk analysis methods, which are heavily dependent on subjective data sources and/or do not consider variable interdependencies within the data, are used. Nonetheless, exploiting the power of ML data analytics tools can result in significant corporate benefit by enhancing the time performance of construction projects—regarded as one of the key indicators of a successful project.

The present study contributed to this endeavor by identifying and applying ML algorithms to develop two construction project delay risk predictive models based on decision tree and naïve Bayesian classification algorithms. This contribution was realized by reaching two key objectives. First, the main influential risk factors and sources affecting construction projects' delays were identified through a literature survey and consultations with construction sector experts. Subsequently, a data set, comprising previous building projects' extents of time overrun and the corresponding contributions of risk sources, was assembled through meetings with construction firms. Throughout this process, several key constraints were considered to ensure data consistency and quality. In addition, and through an exploratory and sensitivity data analysis, an understanding of the complex nature of the construction sector and the interdependence among the various delay risk sources was reached.

Second, a ML-based approach was considered the most suitable to handle such a complex system of interacting variables. Afterward, two different ML algorithms were carefully selected based on the assembled project data's properties and were employed to create trained predictive models. Finally, the models were evaluated using 10-fold cross validation, among other methods, to generate overall and class performance indices through confusion matrices. The results confirmed the validity of both models and the effectiveness of their predictive performance. The analysis further revealed that, based on both training and testing results, the naïve Bayesian model outperforms the decision tree model in terms of overall performance, as well as in every class performance measure. This finding reflected a consensus with the preliminary hypothesis due to the conditional independence of the data variables.

Although the proposed ML analysis approach is thought to be applicable to tackle complex and interdependent systems of risk sources such as those generated within the construction sector, the specific constraints, properties, and limitations associated with the data set analyzed within this study renders the resulting numerical/categorical values not necessarily transferable to other cases/data sets, as is the case of any data-driven model. Nonetheless, the procedures described in the paper can be applied to other project data sets, different from the one compiled herein which was studied mainly to facilitate understanding and demonstrate applicability of the proposed ML analysis approach. Subsequently, some recommendations pertaining to future adoption of the methodology described in the paper are warranted.

First, only variables belonging to the nine identified delay risk sources were considered in the present study. As such, the data set used in the study had multiple constraints on other external project variables to facilitate homogeneity and meaningful analyses. It is recommended, however, that the influences of other external project variables, which were constrained in the present study, are considered in future applications, as project location, type/end use, duration, contract value, contract type, technical complexity, and surrounding area.

Second, the independent variables (nine risk sources) in the data set were found to reflect properties of conditional independence with one another for each class of the dependent variable (TO). Such properties may not be present in other cases and variables may be correlated. It is thus recommended for future studies that careful sensitivity analysis be carried out primarily as a key step on which to base the selection of adequate ML algorithms for application.

Finally, data set size can significantly impact ML model performance results. Small-sized data sets increase the chance of model overfitting, thus adversely affecting model performance. In such cases, the authors highly recommend selecting models suited to small-sized data sets with a demonstrated history of satisfactory performance. On the other hand, larger data sets are more prone to noisy data, which can also undermine model performance. Therefore, noise modeling and outlier analysis techniques are highly endorsed for larger data sets.

Ultimately, the developed methodology can be further incorporated into construction management information systems in support of a proactive project risk management approach that benefits project managers in a twofold manner. The first benefit is the ability to assess and anticipate the time performance of projects, described as the extent of time overrun, from the early planning stages based on the projects' inherent risk levels quantified from these stages. The second benefit pertains to the potential of facilitating continuously refined and more realistic estimates of project durations as the project progresses and while risk factors affecting construction delays are active and dynamic. Overall, such an intelligent platform would influence the state of the practice by addressing the need for transforming multidimensional historical data of completed projects into useful corporate value. Such value would enable construction firms to make knowledge- and evidence-based changes and data-supported decisions to avoid future construction delays.

Data Availability Statement

Data generated or analyzed during the study are available from the corresponding author by request.

Acknowledgments

The financial support for the study was provided through the Natural Science and Engineering Research Council (NSERC) of Canada.

References

Abd El-Razek M. E., H. A. Bassioni, and A. M. Mobarak. 2008. "Causes of delay in building construction projects in Egypt." *J. Constr. Eng. Manage.* 134 (11): 831–841. https://doi.org/10.1061/(ASCE)0733-9364(2008)134:11(831).

Aburrous, M., M. A. Hossain, K. Dahal, and F. Thabtah. 2010. "Predicting phishing websites using classification mining techniques with experimental case studies." In *Proc., Information Technology: New Generations (ITNG), 2010 7th Int. Conf.*, 176–181. New York: IEEE.

Aggarwal, C. C. 2016. *Data mining: The textbook*, 285–426. Berlin: Springer.

Aibinu, A. A., and G. O. Jagboro. 2002. "The effects of construction delays on project delivery in Nigerian construction industry." *Int. J. Project Manage.* 20 (8): 593–599. https://doi.org/10.1016/S0263-7863(02)00028-5.

Al-Momani, A. H. 2000. "Construction delay: A quantitative analysis." *Int. J. Project Manage.* 18 (1): 51–59. https://doi.org/10.1016/S0263-7863(98)00060-X.

Alzahrani, J. I., and M. W. Emsley. 2013. "The impact of contractors' attributes on construction project success: A post construction evaluation." *Int. J. Project Manage.* 31 (2): 313–322. https://doi.org/10.1016/j.ijproman.2012.06.006.

Amor, N. B., S. Benferhat, and Z. Elouedi. 2004. "Naive Bayes vs decision trees in intrusion detection systems." In *Proc., 2004 ACM Symp. on applied computing*, 420–424. New York: Association for Computing Machinery.

Arditi, D., G. T. Akan, and S. Gurdamar. 2006. "Reasons for delays in public projects in Turkey." *Constr. Manage. Econ.* 3 (2): 171–181. https://doi.org/10.1080/01446198500000013.

Arlot, S., and A. Celisse. 2010. "A survey of cross-validation procedures for model selection." *Stat. Surv.* 4: 40–79. https://doi.org/10.1214/09-SS054.

Ashari, A., I. Paryudi, and A. M. Tjoa. 2013. "Performance comparison between naïve Bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool." *Int. J. Adv. Comput. Sci. Appl.* 4 (11): 33–39.

Assaf, S. A., and S. Al-Hejji. 2006. "Causes of delay in large construction projects." *Int. J. Project Manage.* 24 (4): 349–357. https://doi.org/10.1016/j.ijproman.2005.11.010.

Aziz, R. F. 2013. "Ranking of delay factors in construction projects after Egyptian revolution." *Alexandria Eng. J.* 52 (3): 387–406. https://doi.org/10.1016/j.aej.2013.03.002.

Bilal, M., L. O. Oyedele, J. Qadir, K. Munir, S. O. Ajayi, O. O. Akinade, H. A. Owolabi, H. A. Alaka, and M. Pasha. 2016. "Big Data in the construction industry: A review of present status, opportunities, and future trend." *Adv. Eng. Inf.* 30 (3): 500–521. https://doi.org/10.1016/j.aei.2016.07.001.

Breiman, L. 2001. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)." *Stat. Sci.* 16 (3): 199–231. https://doi.org/10.1214/ss/1009213726.

Caldas, C. H., and L. Soibelman. 2003. "Automating hierarchical document classification for construction management information systems." *Autom. Constr.* 12 (4): 395–406. https://doi.org/10.1016/S0926-5805(03)00004-9.

Chan, D. W., and M. M. Kumaraswamy. 1997. "A comparative study of causes of time overruns in Hong Kong construction projects." *Int. J. Project Manage.* 15 (1): 55–63. https://doi.org/10.1016/S0263-7863(96)00039-7.

Chao, L. C., and C. F. Chien. 2009. "Estimating project S-curves using polynomial function and neural networks." *J. Constr. Eng. Manage.* 135 (3): 169–177. https://doi.org/10.1061/(ASCE)0733-9364(2009)135:3(169).

Cheng, M. Y., and Y. W. Wu. 2009. "Evolutionary support vector machine inference system for construction management." *Autom. Constr.* 18 (5): 597–604. https://doi.org/10.1016/j.autcon.2008.12.002.

Cheung, S. O., T. W. Yiu, and H. W. Chan. 2010. "Exploring the potential for predicting project dispute resolution satisfaction using logistic regression." *J. Constr. Eng. Manage.* 136 (5): 508–517. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000157.

Chi, S., S. J. Suk, Y. Kang, and S. P. Mulva. 2012. "Development of a data mining-based analysis framework for multi-attribute construction project information." *Adv. Eng. Inf.* 26 (3): 574–581. https://doi.org/10.1016/j.aei.2012.03.005.

Chou, J. S., and C. Lin. 2013. "Predicting disputes in public-private partnership projects: Classification and ensemble models." *J. Comput. Civ. Eng.* 27 (1): 51–60. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000197.

Desai, V. S., and S. Joshi. 2010. "Application of decision tree technique to analyze construction project data." In *Proc., Int. Conf. on Information Systems*, 304–313. Berlin: Springer.

Dindarloo, S. R., and E. Siami-Irdemoosa. 2016. "Data mining in mining engineering: Results of classification and clustering of shovels failures data." *Int. J. Min. Reclam. Environ.* 31 (2): 105–118. https://doi.org/10.1080/17480930.2015.1123599.

Elazouni, A. M. 2006. "Classifying construction contractors using unsupervised-learning neural networks." *J. Constr. Eng. Manage.* 132 (12): 1242–1253. https://doi.org/10.1061/(ASCE)0733-9364(2006)132:12(1242).

Ferson, S. 2008. "What Monte Carlo methods cannot do." *Hum. Ecol. Risk Assess: Int. J.* 2 (4): 990–1007. https://doi.org/10.1080/10807039609383659.

Flath, C., D. Nicolay, T. Conte, C. van Dinther, and L. Filipova-Neumann. 2012. "Cluster analysis of smart metering data." *Bus. Inf. Syst. Eng.* 4 (1): 31–39. https://doi.org/10.1007/s12599-011-0201-5.

Fugar, F. D., and A. B. Agyakwah-Baah. 2010. "Delays in building construction projects in Ghana." *Constr. Econ. Build.* 10 (1–2): 103–116. https://doi.org/10.5130/AJCEB.v10i1-2.1592.

Gerassis, S., J. E. Martín, J. T. García, A. Saavedra, and J. Taboada. 2017. "Bayesian decision tool for the analysis of occupational accidents in the construction of embankments." *J. Constr. Eng. Manage.* 143 (2): 04016093. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001225.

Glur, C. 2018. "data.tree: General purpose hierarchical data structure: R package version 0.7.6." Accessed August 10, 2018. https://CRAN.R-project.org/package=data.tree.

Goldstein, M. 2006. "Subjective Bayesian analysis: Principles and practice." *Bayesian Anal.* 1 (3): 403–420. https://doi.org/10.1214/06-BA116.

Gong, J., C. H. Caldas, and C. Gordon. 2011. "Learning and classifying actions of construction workers and equipment using Bag-of-Video-Feature-Words and Bayesian network models." *Adv. Eng. Inf.* 25 (4): 771–782. https://doi.org/10.1016/j.aei.2011.06.002.

Guyonnet, D., B. Bourgine, D. Dubois, H. Fargier, B. Come, and J. P. Chilès. 2003. "Hybrid approach for addressing uncertainty in risk assessments." *J. Environ. Eng.* 129 (1): 68–78. https://doi.org/10.1061/(ASCE)0733-9372(2003)129:1(68).

Hastie, T., J. Friedman, and R. Tibshirani. 2009. *The elements of statistical learning*. New York: Springer.

Heravi, G., and E. Eslamdoost. 2015. "Applying artificial neural networks for measuring and predicting construction-labor productivity." *J. Constr. Eng. Manage.* 141 (10): 04015032. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001006.

Hornik, K., C. Buchta, and A. Zeileis. 2009. "Open-source machine learning: R meets Weka." *Comput. Stat.* 24 (2): 225–232. https://doi.org/10.1007/s00180-008-0119-7.

Huang, H. T., and H. P. Tserng. 2018. "A study of integrating support-vector-machine (SVM) model and market-based model in predicting Taiwan construction contractor default." *J. Civ. Eng.* 22 (12): 4750–4759. https://doi.org/10.1007/s12205-017-2129-x.

Hwang, J. S., and Y. S. Kim. 2016. "A bid decision-making model in the initial bidding phase for overseas construction projects." *J. Civ. Eng.* 20 (4): 1189–1200. https://doi.org/10.1007/s12205-015-0760-y.

Ismail, I., A. H. Memon, and I. A. Rahman. 2014. "Expert opinion on risk level for factors affecting time and cost overrun along the project lifecycle in Malaysian construction projects." *Int. J. Constr. Technol. Manage.* 1 (2): 10–15.

Jiang, X., and S. Mahadevan. 2008. "Bayesian probabilistic inference for nonparametric damage detection of structures." *J. Eng. Mech.* 134 (10): 820–831. https://doi.org/10.1061/(ASCE)0733-9399(2008)134:10(820).

Kerzner, H. 2017. *Project management: A systems approach to planning, scheduling, and controlling*. Hoboken, NJ: Wiley.

Kim, H., L. Soibelman, and F. Grobler. 2008. "Factor selection for delay analysis using knowledge discovery in databases." *Autom. Constr.* 17 (5): 550–560. https://doi.org/10.1016/j.autcon.2007.10.001.

Kohavi, R. 1995. "A study of cross-validation and bootstrap for accuracy estimation and model selection." *Ijcai* 14 (2): 1137–1145.

Kokkaew, N., and W. Wipulanusat. 2014. "Completion delay risk management: A dynamic risk insurance approach." *J. Civ. Eng.* 18 (6): 1599–1608. https://doi.org/10.1007/s12205-014-1128-4.

Kuhn, M., and R. Quinlan. 2018. "C50: C5.0 decision trees and rule-based models: R package version 0.1.2." Accessed August 10, 2018. https://CRAN.R-project.org/package=C50.

Lam, K. C., M. C. K. Lam, and D. Wang. 2010. "Efficacy of using support vector machine in a contractor prequalification decision model." *J. Comput. Civ. Eng.* 24 (3): 273–280. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000030.

Mahamid, I. 2011. "Risk matrix for factors affecting time delay in road construction projects: Owners' perspective." *Eng. Constr. Archit. Manage.* 18 (6): 609–617. https://doi.org/10.1108/09699981111180917.

Majid, I. 2006. "Causes and effect of delays in Aceh construction industry." Master of Science thesis, Dept. of Civil Engineering, Univ. Technology Malaysia.

Mansfield, N. R., O. O. Ugwu, and T. Doran. 1994. "Causes of delay and cost overruns in Nigerian construction projects." *Int. J. Project Manage.* 12 (4): 254–260. https://doi.org/10.1016/0263-7863(94)90050-7.

Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. 2017. "Misc functions of the department of statistics, probability theory group (Formerly: E1071), TU Wien. R package version 1.6-8." Accessed August 10, 2018. https://CRAN.R-project.org/package=e1071.

Mezher, T. M., and W. Tawil. 1998. "Causes of delays in the construction industry in Lebanon." *Eng. Constr. Archit. Manage.* 5 (3): 252–260. https://doi.org/10.1108/eb021079.

Ng, A. Y., and M. I. Jordan. 2002. "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes." In *Advances in neural information processing systems*, 841–848. San Mateo, CA: Morgan Kaufmann Publishers.

Niuniu, X., and L. Yuxun. 2010. "Notice of retraction review of decision trees." In Vol. 5 of *Proc., 3rd IEEE Int. Conf. for Computer Science and Information Technology (ICCSIT)*, 105–109. New York: IEEE.

Odeh, A. M., and H. T. Battaineh. 2002. "Causes of construction delay: Traditional contracts." *Int. J. Project Manage.* 20 (1): 67–73. https://doi.org/10.1016/S0263-7863(00)00037-5.

PMI (Project Management Institute). 2017. *A guide to the project management body of knowledge*. 6th ed. Newtown Square, PA: PMI.

R Core Team. 2013. *A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rezaie, K., M. S. Amalnik, A. Gereie, B. Ostadi, and M. Shakhseniaee. 2007. "Using extended Monte Carlo simulation method for the improvement of risk management: Consideration of relationships between uncertainties." *Appl. Math. Comput.* 190 (2): 1492–1501. https://doi.org/10.1016/j.amc.2007.02.038.

Ripley, B. 2018. "Classification and regression trees: R package version 1.0-39." Accessed August 10, 2018. https://CRAN.R-project.org/package=tree.

Sadeghi, N., A. R. Fayek, and W. Pedrycz. 2010. "Fuzzy Monte Carlo simulation and risk assessment in construction." *Comput.-Aided Civ. Infrastruct. Eng.* 25 (4): 238–252. https://doi.org/10.1111/j.1467-8667.2009.00632.x.

Sambasivan, M., and Y. W. Soon. 2007. "Causes and effects of delays in Malaysian construction industry." *Int. J. Project Manage.* 25 (5): 517–526. https://doi.org/10.1016/j.ijproman.2006.11.007.

Seong, H., H. Son, and C. Kim. 2018. "A comparative study of machine learning classification for color-based safety vest detection on construction-site images." *J. Civ. Eng.* 22 (11): 4254–4262. https://doi.org/10.1007/s12205-017-1730-3.

Singh, R. 2009. *Delays and cost overruns in infrastructure projects: An enquiry into extents, causes and remedies*. Helsinki, Finland: Centre for Development Economics.

Son, H., C. Kim, N. Hwang, C. Kim, and Y. Kang. 2014. "Classification of major construction materials in construction environments using ensemble classifiers." *Adv. Eng. Inf.* 28 (1): 1–10. https://doi.org/10.1016/j.aei.2013.10.001.

Therneau, T., and B. Atkinson. 2018. "Recursive partitioning and regression trees: R package version 4.1-13." Accessed August 10, 2018. https://CRAN.R-project.org/package=rpart.

Tixier, A. J. P., M. R. Hallowell, B. Rajagopalan, and D. Bowman. 2017. "Construction safety clash detection: Identifying safety incompatibilities among fundamental attributes using data mining." *Autom. Constr.* 74 (Feb): 39–54. https://doi.org/10.1016/j.autcon.2016.11.001.

Wei, Z., W. Wang, J. Bradfield, J. Li, C. Cardinale, E. Frackelton, and R. N. Baldassano. 2013. "Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease." *Am. J. Hum. Genet.* 92 (6): 1008–1012. https://doi.org/10.1016/j.ajhg.2013.05.002.

Xhemali, D., C. J. Hinde, and R. G. Stone. 2009. "Naive Bayes vs. decision trees vs. neural networks in the classification of training web pages." *Int. J. Comput. Sci.* 4 (1): 16–23.

Xia, N., R. Zhong, C. Wu, X. Wang, and S. Wang. 2017. "Assessment of stakeholder-related risks in construction projects: Integrated analyses of risk attributes and stakeholder influences." *J. Constr. Eng. Manage.* 143 (8): 04017030. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001322.