Predicting the occurrence of construction disputes using machine learning techniques

Article

Accepted Version

Predicting the Occurrence of Construction Disputes Using Machine Learning Techniques

Murat Ayhan, Ph.D.[1]; Irem Dikmen[2]; and M. Talat Birgonul[3]

Abstract

Construction industry is overwhelmed by increasing number and severity of disputes. The primary objective of this research is to predict the occurrence of disputes by utilizing machine learning (ML) techniques on empirical data. For this reason, variables affecting dispute occurrence were identified from the literature and a conceptual model was developed to depict the common factors. Based on the conceptual model, a questionnaire was designed to collect empirical data from experts. Chi-square tests were conducted to reveal the associations between input variables and dispute occurrence. Alternative classification techniques were tested, and support vector machines (SVM) classifier achieved the best average accuracy (90.46%). Ensemble classifiers combining the tested classification techniques were developed for enhanced prediction performance. Experimental results showed that the best ensemble classifier, obtained from majority voting technique, can achieve 91.11% average accuracy. Based on Chi-square tests, the most influential factors on dispute occurrence were found as variations and unexpected events in projects. Other important predictors were all related to skills of the parties involved. This study contributes to the construction dispute domain in three ways (1) by proposing a conceptual model that combined the diverse efforts in the literature for identifying variables affecting dispute occurrence, (2) by highlighting the influential factors such as response rate and communication skills as indicators for potential disputes, (3) by providing an empirical ML-based model with enhanced prediction capabilities that can function as an early-warning mechanism for decision-makers.

Keywords: Dispute prediction; Machine learning; Data classification; Construction disputes; Dispute management; Project management.

Introduction

Numerous parties having different expertise, background, and goals are involved in a construction project in a coordinated manner; however, these parties also have competing goals and expectations as they seek to maximize their own profits simultaneously, which may lead to differences in perception and conflicts in interests (Soni et al. 2017). In addition, construction projects are complicated in nature and their performance is highly susceptible to

[1]Dept. of Civ. Engrg., Gazi Univ., Maltepe, Ankara, 06570, Turkey. Email: muratayhan@gazi.edu.tr
[2] Prof. Dept. of Civ. Engrg., Middle East Tech. Univ., Cankaya, Ankara, 06800, Turkey.
[3] Prof. Dept. of Civ. Engrg., Middle East Tech. Univ., Cankaya, Ankara, 06800, Turkey.

several uncertainties. Stemming from their large and complex nature, construction projects involve large number of uncertainties that makes encountering conflicting situations more common than many other industries (Dalton and Shehadeh 2003). Indeed, there is evidence showing that the construction industry is more problematic compared with other industries (Tazelaar and Snijders 2010). This is mainly due to the fact that it is almost impossible to incorporate provisions to deal with all the possible contingencies due to the large number of uncertainties faced in a typical construction project (Cheung and Pang 2013). Uncertainties may lead to conflicts and when a conflict is not satisfactorily settled, it can quickly escalate to a claim and ultimately a dispute. With a potential to result in delayed schedules, budget overruns, poor quality and performance, increased tension, and damaged long-term business relationships, construction disputes can be detrimental (Cheung and Suen 2002).

The current tendency in the construction industry is to make challenging decisions related to dispute management actions intuitively based on the experience of the decision-maker. The complexity of decision-making increases considering that the availability of the information is limited, and its quality is questionable (Chou et al. 2013b). Hence, experience and knowledge are invaluable in decision-making (Cheung et al. 2004; Mokhtar and Rahman 2017). The merits of Artificial Intelligence (AI) techniques include extraction of tacit knowledge in an articulable and presentable way to the decision-makers. In computer science, AI involves the systems created to perform tasks that usually require human intelligence and among various AI applications, data mining (DM) via machine learning (ML) techniques form an important research branch as they enable gathering valuable information from large volumes of data that is difficult to understand and interpret (Liao et al. 2012). In DM, which is a subset of AI, large volumes of data are processed to establish simple models with valuable use that enable identification of hidden knowledge and patterns in the data (Bilal et al. 2016). ML is a subset of DM that resorts to statistical theories for building models (i.e., predictive) and it includes the algorithms that help machines to learn from past data systematically and automatically based on optimization of a performance criterion (Alpaydin 2010; Bilal et al. 2016). As being dependent to experience and knowledge of the decision-maker rather than being systematic, the current decision-making practice in dispute management is prone to subjectivity (Parikh et al. 2019). AI applications, on the other hand, has the potential to minimize this subjectivity by providing systematical decision-support based on past cases that fits the circumstances of the current case (Cheung et al. 2004).

This paper argues that in order to forestall construction disputes, prediction models for dispute occurrence could be developed to minimize the subjectivity of the decisions made by the decision-makers. To this end, the primary objective of this paper is to develop a new construction dispute prediction model by utilizing ML techniques on empirical data. The proposed model is expected to create an early-warning mechanism that will enable taking

informed management actions. Thus, the expected outcome of this research is to help construction professionals in taking the necessary precautions against disputes by early-warning.

In accordance with the research objectives, in this paper, first, research background and findings of literature review are given, and research methodology is explained. Variables affecting dispute occurrence identified from the literature are discussed and a conceptual model is presented to depict the common factors that relate to dispute occurrence. Then, the questionnaire to collect empirical data from decision-making authorities is discussed and results of Chi-square tests are presented. Chi-square results were also validated by measuring the information gain ratio in order to evaluate the worth of each individual input variable with respect to dispute occurrence. Finally, finalized prediction model that was tested by using alternative single and ensemble ML techniques to obtain the best classifiers is given. The findings are discussed along with limitations of the study and it is followed by concluding remarks and recommendations for future studies.

Research Background and Motivation

The severity of disputes in construction has been well understood and documented; however, the construction industry still struggles to find methods to resolve them fairly and economically (Cheng et al. 2009). According to annual reports of Hong Kong International Arbitration Centre (HKIAC), the average rate of construction disputes was 20.2% among all HKIAC registered cases between 2015 and 2017 (HKIAC 2018). The American Arbitration Association (AAA) reported that the number of construction cases submitted to the AAA in 2017 were up by 4%, which involved 13% increase in claims higher than 1 million U.S. Dollars (AAA, 2018). Awwad et al. (2016) stated that the growth of the construction industry in the Middle East region is accompanied by an increasing number of construction disputes and growing number of arbitration cases are being witnessed. The study by Parikh et al. (2019) revealed that the National Highways Authority of India (NHAI) is struggling with more than 1000 construction disputes amounting over 1 billion Indian Rupees and meanwhile, the occurrence of claims and disputes are on the rise among over 200 contracts under implementation. Ustuner and Tas (2019) drew attention to huge numbers of disputed cases submitted to resolution organizations. In their study, it is reported that the Judicial Arbitration and Mediation Services (JAMS), which is among the largest dispute resolution organizations in the world, handles approximately 15000 cases annually; and the Centre for Effective Dispute Resolution (CEDR), which is a U.K. based organization, handles approximately 30000 disputes annually. Considering all these cases, it can be concluded that the disputed projects constitute a significant portion of the construction industry and that an incremental trend is observed in the number of disputes in many regions of the world. The financial consequences of disputes are also significant. The estimated additional direct costs of disputes range from

0.5% to 5% of the contract value (Love et al. 2010). On the other hand, there are additional indirect costs due to decreased productivity, strained business relationships, loss of future business opportunities, and damaged reputation of parties that amplify the damages caused by disputes (Ilter 2012). Supported by the above-mentioned examples, it is no surprise that construction disputes drew attention of many researchers.

Diekmann and Girard (1995) conducted one of the pioneering studies on predicting the occurrence of disputes. They developed a 'dispute potential index (DPI)' using Logistic Regression (LR) on a dataset of 159 construction projects. An improvement was achieved by analyzing the same dataset using the Structural Equation Modeling (SEM) (Molenaar et al. 2000). Both studies aimed to predict dispute propensity at early stages. However, construction disputes require consideration of numerous complex and interrelated factors that are difficult to rationalize (Chou 2012). Therefore, results from techniques like LR and SEM, which have limitations in describing nonlinear relationships (An et al. 2007) and in modeling multiple correlations, can be misguiding. In this paper, these limitations are addressed by ML techniques that can reflect the complex interrelationships between variables. In another notable study, Cheung and Pang (2013) proposed an anatomy of construction disputes arranged under fault-tree methodology that identified various causes of disputes, categorized them under adequate factor groups, and assessed the fuzzy occurrence likelihood of these factors, which gradually led to dispute occurrence likelihood evaluation. Although the anatomy was obtained from an expert panel, it was not supported by empirical data from real-world construction projects. Indeed, researchers highlighted the lack of empirical evidence in the construction disputes literature to support the presented theories (Love et al. 2010; Ilter 2012). Being one of the concerns of the presented paper, this gap is addressed by supporting the proposed prediction model with empirical data.

It is commonly accepted that the best solution against disputes is to avoid them and the actions for avoidance can only be taken by prediction (Fenn 2007). Considering the fact that developing deterministic mathematical models to solve construction management and prediction problems is difficult and costly, the research interest moves towards approximate inference as a fast and cost efficient alternative and consequently, the use of AI is appropriate in efforts to solve such problems (Cheng and Wu 2009). Among various AI applications, the use of DM techniques is accepted as a strongly effective method for determining numerous complex and interconnected factors related to construction disputes along with hidden relationships that are difficult to rationalize (Chou 2012). For example, it might be intuitively expected that with an increasing level of construction complexity, more disputes would occur. However, the experience of the contractor on the type of project may alter the relationship between construction complexity and dispute occurrence. Considering solely the relationship between the input variable (i.e., level of construction complexity) and the output (i.e., dispute occurrence) may not reveal the

complete picture and may lead to misinterpretation of the circumstances. DM techniques are well-suited to determine such relationships. Especially, algorithms available in the ML domain (i.e., Decision Trees (DT), Support Vector Machines (SVM)) makes developing data-specific prediction models possible. When the output variable is a categorical variable, prediction problems become data classification problems (Chou and Lin 2013). Data classification problems are problems of associating an instance (a case), which is defined by values of its attributes (observed variables), with a class among predefined classes as an output (Pulket and Arditi 2009). In the case of predicting the occurrence of disputes, a construction project is an instance, various characteristics related to a construction project are its attributes and the output variable is the occurrence of disputes, where projects can be categorized as 'disputed' and 'undisputed' projects. Using ML techniques, this paper aims to classify future construction projects as disputed or undisputed based on empirical data collected from past construction projects.

According to a study published in 2009, the tendency in disputes related literature is to produce general insights and statistical outcomes rather than establishing supporting models or systems (Ilter and Dikbas 2009). After more than a decade, the number of studies that establishes models or systems are still limited and there are even less studies that assess the dispute propensity of construction projects by utilizing ML techniques. Moreover, these limited studies suffer from various shortcomings. Existing studies are mainly specific to a certain project type (i.e., Public-Private-Partnership (PPP) projects) (Chou and Lin 2013; Chou et al. 2013a), or to a certain dispute type (i.e., change order related disputes) (Chen and Hsu 2007), or to cases from a certain country or region (Yousefi et al. 2016). The proposed study addresses these needs systematically by collecting data of various construction projects from different regions of the world. The dataset is not limited with a single project or dispute type.

A series of studies on a dispute dataset of PPP projects undertaken in Taiwan proved the effectiveness of ML techniques in dispute occurrence prediction (Chou et al.2013a; Chou and Lin 2013; Chou et al. 2014; Chou et al. 2016). The first of these studies (Chou et al. 2013a) utilized k-Nearest Neighbor (kNN), Multilayer Perceptron (MLP), Naïve Bayes, SVM, and C4.5 as single classifiers. Then, in pursuit of enhanced classification performance, ensemble classifiers were established by combining k-means Clustering technique, MLP classifier, and C4.5 algorithm, respectively, with the mentioned single classifiers one by one. It was highlighted that the prediction performance of ensemble models outperformed the single classifiers. In the second study (Chou and Lin 2013), the same dataset was analyzed by several single and ensemble classifiers again, and according to 10-fold cross-validation (CV) performances, the highest prediction rate was 84.33%. The third study (Chou et al. 2014) focused on SVM algorithm in specific to achieve 89.30% accuracy. Finally, in the fourth study (Chou et al. 2016), the dispute occurrence was predicted by C5.0 algorithm with an average 10-fold CV accuracy of 83.92%.

Similar to predicting the occurrence of disputes, there are studies predicting the litigation likelihood. A noted study was based on a dataset of 340 litigated cases with change order related problems in the U.S. Using this dataset, a hybrid model was developed by combining Artificial Neural Network (ANN) and Case-Based Reasoning (CBR) techniques to classify construction projects according to their litigation likelihood. The model achieved 84.61% accuracy (Chen and Hsu 2007). However, the dataset was composed of litigious cases only, which ignores valuable knowledge that could have come from cases solved by other resolution methods. Moreover, this research was specific to the U.S. construction industry and change order related disputes. In their prediction models, Mahfouz et al. (2018) provided legal decision support by enabling automatic extraction of implicit knowledge about significant factors upon which verdicts of differing site condition (DSC) litigations are based. Their study was based on 600 cases from the Federal Court of New York and several prediction models were developed using ML techniques. However, the study was limited to DSC litigations from New York only.

In the light of foregoing observations, it can be concluded that the subject matter of existing studies on dispute prediction have generally focused on specific points (i.e., prediction of change order related disputes or disputes in a certain region). There is a dearth of models that incorporate various project, dispute, and project delivery system types in the literature. Besides the main goal of the paper to develop a new model for predicting the occurrence of construction disputes, another goal is related to address the lack of empirical evidence in the construction disputes literature by using a more generic dataset that reflects variations in construction types, project delivery systems, etc. while achieving an advancement in the accuracy of predictions compared to previous studies.

Research Methodology

The research design includes three steps. As it is presented in Fig. 1, the first step involves the conceptual model development, the second step is the development of the prediction model, and the third step is finalization of the prediction model.

Conceptual Model Development

There is a diversity in the literature about the causes of disputes or factors that affect them. Moreover, there is a confusion in the related terminology due to overlapping concepts; the distinction between causes, factors, or types of disputes are rather vague (Ilter 2012). In order to overcome these problems, an extensive literature review on conflicts, claims, and disputes was conducted with the aim of synthesizing the findings of the previous research in a conceptual model. To understand mechanisms of dispute development, researchers tried to identify causes of disputes and project characteristics or attributes that impact their occurrence. Besides, there are many other studies with various goals such as predicting the dispute proneness of projects (Diekmann and Girard 1995; Chou and Lin

2013; Chou et al. 2013a; Chou et al. 2014; Yousefi et al. 2016) that identify dispute causes or impacting attributes not as a primary goal, but as a secondary goal. Another group of studies can be found in research aiming to analyze the outcome of claims, disputes, and resolution methods (i.e., litigation outcome). Specifically, there are studies focusing on dispute causes and project or contract characteristics that affect court rulings (Kilian and Gibson 2005; Chen and Hsu 2007; Pulket and Arditi 2009; Arditi and Pulket 2010). Within this context, research in the above-mentioned fields are reviewed within reputable journals from 1995 to 2018 including Journal of Construction Engineering and Management, ASCE; Journal of Computing in Civil Engineering, ASCE; Journal of Civil Engineering and Management; International Journal of Project Management; and Expert Systems with Applications. Numerous causes and reasons of disputes along with project characteristics were reviewed, and frequently perceived parameters were taken as the most prominent factors. The findings of the review showed that the prominent factors that impact dispute occurrence are related to (1) characteristics of the project (i.e., duration, value), (2) characteristics of the parties involved in construction projects and their organizational structures (i.e. response rate and communication skills), (3) occurrence of changes or unexpected events, and (4) delays. These categories were established by combining the findings of several research including (1) Diekmann and Girard (1995) and Molenaar et al. (2000) that grouped their attributes under project, people, and process related aspects, (2) Chen and Hsu (2007) that grouped attributes related to project characteristics and delays under project data category, and attributes related to changes and other dispute characteristics under disputed issue data category, (3) Dalton and Shehadeh (2003) that focused on attributes related to skills of the parties, (4) Ilter and Dikbas (2009) and Ilter (2012) where a set of variables were designed to provide data on several project characteristics as well as project managers and their firms. These categories were used in the conceptual model to group the identified input variables. The detailed explanation related to determination of input variables from the literature is given in Ayhan (2019) that developed a conceptual model to depict the common factors influencing dispute occurrence, compensations related to disputes, and the resolution method selection. Based on the conceptual model, empirical data was collected with the aim of developing prediction models. The developed prediction models classified the dispute occurrence of construction projects; then, among the disputed projects, the type of compensation related to the dispute was predicted; and finally, the adequate resolution method was predicted. The proposed paper here reflects the first step in Ayhan (2019)'s study that presented a prediction model for dispute occurrence.

The first category in the conceptual model is named as 'Project Characteristics (PC)' that involves 11 attributes related to project and contract related characteristics of a construction project. The second category is the 'Skills (S)' that is composed of 12 attributes depending on parties involved in the project and their organizational

structures. The third category of input variables is the 'Changes (C)' category that involves occurrence of variations and unexpected events in a construction project. Finally, the fourth category is the 'Delays (D)' that considers the impact of delays on construction disputes.

Development of the Prediction Model

The second step is the development of the prediction model for dispute occurrence. The model relies on empirical data from past construction projects. In order to collect construction project data, a questionnaire was designed based on the identified variables in the conceptual model. The first section of the questionnaire is used to collect background information of the participants. The second section was designed to collect information related to project characteristics. Other than the two questions related to level of design and construction complexity, which were measured on a 5-point Likert scale (1: lowest level of complexity, 2: low complexity, 3: moderate complexity, 4: high complexity, 5: highest level of complexity), the remaining questions aimed to gather quantitative data related to the project such as contract value, time extensions, etc. The third section was composed of questions to assess characteristics of the parties involved and their organizational structures (i.e., coordination skills). These were also qualitatively measured by using 5-point Likert scale. The final section was composed of yes/no questions to understand occurrence of changes or unexpected events. The complete version of the questionnaire is available in Ayhan (2019).

In the collected data, the impacts of input variables on the output will not be the same. Some variables may impact the outcome more than the others, while the impact of some can be statistically insignificant. Therefore, the significance of associations between inputs and occurrence of disputes should be analyzed. First, the dataset was cleaned from noisy data and processed such that numeric variables were turned into categorical variables (i.e., nominal and ordinal). Then, to understand if there exists a statistically significant relationship between inputs and the output, Chi-square statistics was utilized. Chi-square statistics is a useful way of testing the existence of association relationship between categorical variables (Weisburd and Britt 2007). The Chi-square tests were performed in IBM SPSS Statistics version 22.0. Finally, according to the results of the Chi-square tests, statistically insignificant variables were eliminated, and prediction model was developed including only the significant variables. In other words, attribute elimination was performed on variables of the conceptual model via Chi-square tests to construct the prediction model for dispute occurrence.

Finalization of the Prediction Model

The third step is finalizing the prediction model via data classification. For this reason, classification performances of alternative single and ensemble ML techniques were experimented. The utilized single algorithms

were (1) Naïve Bayes, (2) kNN, (3) C4.5, (4) MLP, (5) Polynomial kernel SVM, and (6) Radial Basis Function (RBF) kernel SVM. Meanwhile, the ensemble classifiers were developed by using (i) voting technique, (ii) stacked generalization, and (iii) the AdaBoost algorithm. The classifier with the best classification performance was presented as the final prediction model for dispute occurrence. Containing numerous inbuilt ML algorithms, WEKA version 3.8.3 (Frank et al. 2016) was used in data classification experiments. WEKA is a tried and tested software with evidence from the literature showing that the tool has consistent results and comparable performance to other applications, if not superior (Al-Khoder et al. 2015; Arasu et al. 2020). The algorithm parameters in WEKA can be edited easily via user interface, which helps the user to easily enhance the performance of the classification algorithm, instead of having to deal with complex command line operations (Witten et al. 2016).

Data Collection and Initial Findings

With the goal of reflecting varying characteristics, the data was collected from a wide variety of construction projects. The study dataset contained 151 construction projects initially. However, noisy and unrepresentative data were removed before analysis (data cleansing). The remaining dataset for this research involves 108 projects, which are executed in 19 different countries. These projects were obtained from 75 construction companies of six different nationalities via face-to-face and online meetings with 78 individuals. Among these 75 companies, 21.3% are placed in the Engineering News-Record (ENR) Top 250 International Contractors List in 2018 (ENR 2018). Among these 108 projects, 38 of them did not experience any disputes (35%), while 70 projects faced with at least one dispute. This shows that dispute occurrence in construction projects are dominant (65%) for this dataset.

The data was collected from participants having a wide variety of roles including owners (26.9%), project directors (15.7%), legal counselors / advisors (15.7%), contract managers (7.4%), claim / dispute managers (6.5%), project managers (10.2%), site managers (10.2%), and project engineers (7.4%). Moreover, participants were selected with different levels of experience, ranging from 2 years to 49 years. The average construction experience of participants in the dataset is approximately 18 years with 47% having worked more than 15 years. Thus, it can be claimed that mainly the opinions of senior professionals are reflected in this research.

Table 1 shows the categorical labels and frequencies of ‘PC’ attributes in the dataset. It should be noted that numeric attributes are converted into categorical variables (data transformation) for computational purposes. For example, Naïve Bayes classification requires discrete data. However, if the discretization process removes distinguishing features, the data type conversion may harm the accuracy of classification algorithms later. For this reason, this study used the information gain-based supervised discretization available in WEKA that minimizes the subjectivity by considering the output class while determining discretization boundaries, and that minimizes

the information loss by selecting the split points generating the largest information gain. Table 2 shows categorical levels and frequencies of 'S' attributes in the dataset. Table 3 is for 'C' attribute that is measured on a binary scale, while Table 4 holds the information for 'D' attribute. The ratio of extensions to total planned duration is considered as a measure of delay in this paper. However, this numeric attribute is also converted to a categorical variable.

Some attributes reveal their pattern in the dataset at the first glance. For example, the increase in project values causes an increase in dispute rates as shown in Fig. 2. The highest dispute rate is observed for the group of projects with the highest project values (86%), while projects with the lowest values have the lowest dispute rate (48%). Fig. 3 shows the dispute rates with respect to planned project duration. For this dataset, dispute rates indicate that the longer the project duration is, the higher is the dispute rate. Fig. 4 presents the dispute rates with respect to project location, which shows that encountering disputes in international projects (82%) is more likely compared with domestic projects (59%). Thus, construction professionals should pay more attention to dispute management in international construction projects. Finally, the delays attribute showed that when the project is finished without any extensions, the dispute rate is 47.7%. Meanwhile, this rate is 62.5% in projects with an extension to planned duration ratio up to 20%, and 94.1% in projects with an extension ratio between 20% and 40%. These findings are in line with Ilter (2012) that reviewed the literature to identify factors affecting disputes; and project value, duration, delays, and unfamiliarity with local conditions were reported among dispute factors. However, not all relationships are easy to interpret and, attributes can also have a combined effect on the output. Therefore, the dataset is analyzed by the Chi-square tests and ML techniques to reveal all undiscovered associations.

Chi-square Tests on the Dataset and Prediction Model for Occurrence of Disputes

The performance of ML algorithms is generally affected negatively by irrelevant attributes. Therefore, elimination of insignificant attributes improves generalization performance of ML algorithms (Arditi and Pulket 2010; Sonmez and Sozgen 2017). Among numerous attribute elimination methods in the literature, Chi-square statistics is preferred in this paper. It is a non-parametric method that is robust to distribution of the data and unequal variances among study groups (Weisburd and Britt 2007). This means the Chi-square results can compensate the problematic issues due to data distribution (i.e., skewed data) unlike many other methods that require data with almost normal distribution and equality of variances. In addition, this method can handle both dichotomous and multiple category variables (McHugh 2013). Considering that the dataset in this research is composed of dichotomous and multiple category input variables with various distributions, Chi-square statistics is found to be an appropriate evaluation technique for attribute elimination and thus is used in this study.

There are two types of categorical variables in the dataset as nominal and ordinal variables. The information related to the ordering is important for ordinals, and thus, methods for nominals and ordinals should not be applied interchangeably (Agresti 2007). Therefore, in the Chi-square tests, these variables should be analyzed accordingly.

Chi-square tests do not reveal the strength of association between variables (Agresti 2007); consequently, it should be followed with a statistic showing the strength (McHugh 2013). Phi and Cramer's V are two of the measures for evaluating the strength of association between nominal variables. Phi measure calculates a strength value only for 2×2 tables. However, the real-world data and variables generally do not suit this kind of limitation. For this reason, there is another measure called Cramer's V, which has the capability of handling tables with varying numbers of rows and columns (Weisburd and Britt 2007). Therefore, although Cramer's V can calculate low correlation values for highly significant results, it has become the most preferred strength test for nominal variables (McHugh 2013). Consequently, to handle the changing number of rows and columns between variables of this study's dataset, Cramer's V is preferred as a strength measure for nominal variables.

IBM SPSS Statistics version 22.0 presents four measures of association for ordinal variables as Gamma, Somers' d, Kendall's tau-b, and Kendall's tau-c. Strength values obtained from these measures are generally different from each other due to differences in handling tied pairs of observations. In Gamma measure, tied pairs of observations are not considered and consequently, there is a problem of overestimating the strength of the relationships. On the other hand, Somers' d and Kendall's tau measures take tied pairs into account and thus, they are superior to Gamma measure. However, Kendall's tau-b is more adequate when the number of categories of independent and dependent variables are equal. For unequal number of rows and columns, Kendall's tau-c is more adequate. Similar to Kendall's tau-c, Somers' d is also suitable for data with unequal number of rows and columns. Though, it can be considered as a better measure compared with Kendall's tau-c (Weisburd and Britt 2007). In the light of all the facts mentioned above, Somers' d is preferred as a strength measure for ordinal variables.

Chi-square results are tabulated in Table 5 to 8. In these tables, exact probability values are obtained from the exact Pearson Chi-square statistics for nominal variables and the Mantel-Haenszel linear-by-linear association test for ordinals. Probability values are compared with alpha level at '0.05' for 95% confidence interval (CI). Insignificant attributes have an alpha level greater than '0.05'. The significant attributes are in moderately strong, or strong, or very strong relationship with dispute occurrence. When Cramer's V measure is used, 'strong' relationship refers to a value greater than '0.15' and 'very strong' relationship refers to a value greater than '0.25' (Akoglu 2018). In this study, Cramer's V measure of strength is used for only two of the significant attributes, PC1 and C1. Attribute PC1 (Cramer's V = 0.215) has a strong relationship with dispute occurrence, while C1

(Cramer's V = 0.576) has a very strong relationship. Somers' d measure is a proportional error reduction (PRE) measure of association and for PRE values, as a rule of thumb, a value greater than '0.1' refers to a 'moderately strong' relationship and a value greater than '0.4' refers to a 'strong' relationship (Pollock 2011). Among attributes where Somers' d is used, PC2, PC3, S2, S4, S5-1, S5-2, S6-2, S7-2, S8-1, S8-2, and D1 have moderately strong relationships with dispute occurrence, while S1 has strong relationship. In summary, the conceptual model for dispute occurrence prediction had 25 input variables. However, the association is statistically significant in only 14 of them as indicated in Tables 5-8. Fig. 5 shows the established prediction model with these 14 attributes.

Chi-square results are also validated by measuring the information gain ratio, which evaluates the worth of each input variable with respect to the output. WEKA is used to measure the gain ratio using 'GainRatioAttributeEval' as the evaluator. The information gain ratio of each attribute in the conceptual model is measured and ranked. The results are given in Table 9. It is observed that the selected 14 attributes based on Chi-square results are also generating the highest 14 gain ratio values. Thus, Chi-square results are verified by gain ratio values.

ML Techniques

The primary goal of this paper in utilizing ML techniques is to determine patterns in the data so that classifiers can be developed for predicting future cases. For prediction problems, researchers can resort to regression or classification techniques in the supervised ML domain. When the output variable is continuous or numerical, regression techniques can be applied. When the output is categorical, classification techniques should be used. In the case of predicting dispute occurrence, there are two classes that the instances can be assigned, and this kind of classification is called binary classification. ML techniques are well equipped to solve such problems. However, it is difficult to select the best performing ML technique that suits the prediction problem at hand. The literature has proven that it is not possible to solve all DM problems using a single ML technique because of the varying characteristics of real-world datasets (Pulket and Arditi 2009). Instead, to obtain accurate results, the bias due to the learning technique should be compatible with the dynamics of the problem domain, which makes DM an experimental process (Witten et al. 2016). The conventional approach is to experimentally compare performances of promising single ML algorithms as base classifiers and select the best one in that dataset (Arditi and Pulket 2010). For this reason, this paper assessed the performance of six single ML techniques as potential base classifiers.

The reviewed single ML algorithms for data classification are taken from Witten et al. (2016) that listed the top 10 DM algorithms based on results of a poll. These algorithms involve techniques for data classification, clustering, etc. However, proposed dispute prediction model requires utilization of classification techniques. Among these 10 algorithms, the ones that can be used in data classification are tested on the collected data. Besides these algorithms,

MLP is also included because of its common usage in construction research. In short, the six data classification algorithms used in this paper are (1) Naïve Bayes, (2) kNN, (3) DT (C4.5), (4) MLP, (5) polynomial kernel SVM, and (6) RBF kernel SVM. The freely available C4.5 algorithm is preferred in this paper instead of the enhanced C5.0 that is the commercial version of its predecessor. Moreover, there is evidence showing the open-source C4.5 can generate superior or comparable performances to C5.0 (Hssina et al. 2014; Febriantono et al. 2020).

The classification performances of single ML techniques can be enhanced further by creating ensemble classification schemes systematically (Arditi and Pulket 2010). Ensemble approaches, which are simply adding or combining base classifiers, can compensate errors of single classifiers and improve the classification accuracy. There are various approaches to develop ensemble models and this paper utilized three of them, which are (i) voting technique, (ii) stacked generalization, and (iii) the AdaBoost algorithm. This research preferred using the voting technique to benefit from its simplicity since it is reported as the simplest way to combine multiple classifiers (Alpaydin 2010). However, problems of misclassification may occur in voting when majority of the classifiers misclassifies an instance, and it is not clear which classifier's classification is reliable. Unlike the voting technique, in stacked generalization, the classifier to be trusted can be learned by using another classifier (meta-learner) and consequently, this technique may generate better classifier combinations (Witten et al. 2016). The ensemble techniques are used mainly to decrease variance (i.e., bagging), decrease bias (i.e., boosting), and improve prediction accuracy (i.e., stacked generalization). When bagging is used, generating complementary base-learners is by chance; while in boosting, the next classifier is trained systematically on the mistakes of the previous one (Alpaydin 2010). Therefore, this research preferred boosting, specifically the AdaBoost algorithm due to its wide use, ease of implementation, and adaptability to a wide range of classifiers (Witten et al. 2016).

Theoretical background related to tested algorithms will exceed the scope and readers seeking more information can resort to Alpaydin (2010). Each technique has its own parameters that impact the performance of the classifier differently. Table 10 shows parameter configurations for each technique with corresponding search ranges and settings. The optimum values for parameters with numeric search ranges should be determined on a test set to match the characteristics of the dataset, and this is normally determined by using a validation set or CV technique. WEKA provides several methods (i.e., CV parameter selection, grid search) to obtain optimized parameters.

Data Classification Tests using ML Techniques

Model Validation and Performance Evaluation Measures

When limited amount of data is available, splitting the dataset into training and test sets may cause loss of information. Instead, researchers may prefer using all the data for knowledge extraction. However, this leaves no

389 unseen instances, or in other words, test set for the trained classifiers. In such cases, CV technique can be used,
390 which is a procedure using all the data for learning and estimating the accuracy of the classifier by resampling the
391 dataset (Vanwinckelen and Blockeel 2012). K-fold CV is the most common resampling technique that is based on
392 training and testing the model k-times randomly on different subsets of training data to generate an estimate of the
393 performance of a classifier on new data. The k number in k-fold CV is typically '10' (Alpaydin 2010). Although
394 this number can be adjusted depending on the size of the dataset and the desired level of analysis, literature has
395 proven that 10-folds is the right number of folds based on experiments using various datasets and algorithms (Chou
396 and Lin 2013, Chou et al. 2013a; Witten et al. 2016; Sonmez and Sozgen 2017).

397 There are two shortcomings of the k-fold CV technique. Firstly, training and test sets should be representative
398 of the dataset, and random sampling may cause uneven representation. To overcome this problem, there is a
399 procedure called stratification that guarantees each class is properly represented in both training and test sets during
400 random sampling (Witten et al. 2016). Secondly, if two different k-fold CVs are performed using the same
401 algorithm and dataset, but with different random sampling, there will most likely be two quite different
402 classification performances. This is due to high variance associated with CV results. This variance can be restored
403 by repeating the process with different random samples of the same dataset (repeated CV) and taking the average
404 of results obtained from each CV (Vanwinckelen and Blockeel 2012). In the light of these facts, this paper utilized
405 10 times repeated 10-fold CV in evaluating the classifier performance and results are given within 95% CI.

406 In ML domain, confusion matrices are used for evaluating the performances of the classifiers (Sonmez and
407 Sozgen 2017). They are useful in calculating the accuracy, true positive (TP) rate, true negative (TN) rate, false
408 positive (FP) rate, and false negative (FN) rate. Table 11 is a typical confusion matrix. In the ideal case, the
409 diagonal elements, which correspond to correct classifications, should be large and non-diagonal elements, which
410 correspond to misclassifications, should be low (Witten et al. 2016).

411 Accuracy in Eq. 1 is the overall classification performance, and it is the percentage of instances predicted
412 correctly divided by the total number of instances. Precision in Eq. 2 is the number of correctly classified positive
413 instances divided by the number of instances predicted as positive, and it gives the positive predictive power of a
414 classifier. Recall (or sensitivity) in Eq. 3 is the number of correctly classified positive instances divided by the
415 number of positive instances, and it gives the TP rate of a classifier. Specificity in Eq. 4 is the number of correctly
416 classified negative instances divided by the number of negative instances, and it gives the TN rate of a classifier.

$$\text{Accuracy (\%)} = \left(\frac{TP + TN}{TP + TN + FP + FN}\right) 100 \quad (1)$$

417

$$\text{Precision} = \left(\frac{TP}{TP + FP}\right) \quad (2)$$

$$\text{Recall (TP rate)} = \left(\frac{TP}{TP + FN}\right) \quad (3)$$

$$\text{Specificity (TN rate)} = \left(\frac{TN}{TN + FP}\right) = 1 - \text{FP rate} \quad (4)$$

There is a combined measure called the receiver operating characteristic (ROC) curve that characterizes the trade-off between TP and FP rates. The ROC curve depicts the performance of a classifier regardless of class distributions, and it is plotted with the TP rate (recall) on the vertical and the FP rate on the horizontal axis (Alpaydin 2010). The ROC curve indicates the ability of a classifier to avoid misclassifications (Chou and Lin 2013). ROC curves can be summarized in a single quantity that is called the area under the ROC (AUROC), and the larger the area or the closer the AUROC value is to '1', the better the model would be (Witten et al. 2016). The final measure is the Cohen's Kappa coefficient that is used to measure the agreement between predicted and actual values in a dataset with a correction for agreements by chance. Among positive values of the Kappa statistic, Cohen suggested that results can be interpreted as values between (1) 0.01-0.20 indicating none to slight agreement, (2) 0.21-0.40 indicating fair agreement, (3) 0.41-0.60 indicating moderate agreement, (4) 0.61-0.80 indicating substantial agreement, and (5) 0.81-1.00 indicating almost perfect agreement (McHugh 2012).

Results of Data Classification Tests

Table 12 shows the 10 times repeated 10-fold CV results of the single classifiers. The most successful models in terms of average accuracy are RBF kernel SVM, polynomial kernel SVM, and C4.5 classifiers, respectively. The best average Kappa statistic value (0.790), which indicates a substantial agreement between predicted and actual values in the dataset, is obtained from the RBF kernel SVM. The best average precision and specificity values are obtained from the kNN technique, while the RBF kernel SVM generated comparable performances in these measures. The best classifiers in classifying disputed projects (TP rate) are SVM classifiers with an average recall value of 0.929. The best average AUROC value is obtained from the Naïve Bayes classifiers (0.953), which is an almost ideal value. All algorithms produced impressing AUROC values with the lowest value being 0.887.

The best performing three single techniques (RBF kernel SVM, polynomial kernel SVM, and C4.5) are used as candidates for ensemble models. In voting technique, the ensemble classifier will be the combination of these three classifiers. In stacked generalization, two single techniques are combined as base-learner and meta-learner, where base-learner is trained with the original dataset and meta-learner is trained with a subset of the original set that only includes the correctly classified instances by the base-learner. The resulting stacked classifier, which combines performances of two classifiers, is expected to achieve better performance than each single technique it

involves. However, classifiers of the same type should not be combined in stacking (Alpaydin 2010). In summary, for stacked generalization, the best performing three single techniques will be used as base-learners, while all six experimented single techniques will be considered as meta-learners. Keeping in mind that same classifiers should not be combined (i.e., combining two C4.5 classifiers is not desired), the three classifiers that will be used as base-learners can be combined with the remaining five techniques excluding itself. Consequently, this will generate 15 stacked classifiers. Meanwhile, since the main principle of the AdaBoost algorithm is to develop strong classifiers out of weak ones, all single techniques are boosted by the AdaBoost algorithm in pursue of such an enhancement.

Table 13 shows the 10 times repeated 10-fold CV results of the ensemble classifiers. The ensemble classifiers obtained from the majority voting technique generated 91.11% average prediction accuracy. Although the stacked generalization technique that combined RBF and polynomial kernel SVM classifiers also achieved the same average accuracy, the majority voting technique has higher performance as it is superior to all single and ensemble classifiers in terms of average Kappa (0.806), precision (0.937), and specificity (0.884) values. On the other hand, the stacked classifier generated the highest TP rate (0.931) that indicates the success in correct identification of the disputed cases among actually disputed projects, while the majority voting technique has a slightly lower TP rate (0.926). However, when AUROC values are evaluated, the ability to avoid any misclassifications (disputed or undisputed) is better in majority voting (AUROC = 0.905) compared with stacking (AUROC = 0.903).

In stacked generalization, when the classification accuracies of single classifiers contained in the ensemble model is as high as possible and classifiers are as diverse as possible, the ensemble model can outperform the performances of the classifiers it contains (Alpaydin 2010). However, experiments showed that not all 15 stacked ensemble models achieved better classification. When the base-learner is the polynomial kernel SVM or C4.5 algorithm, none of the ensemble models achieved better performances. Moreover, when the base-learner was the RBF kernel SVM, all ensemble models gave the exact same accuracy values. This is because the base-learner does most of the work and the meta-learner is like an arbiter in the stacked ensemble models (Witten et al. 2016). Among the five stacked classifiers where RBF kernel SVM is the base-learner, only the results of the model that combined RBF and polynomial kernel SVMs are given since it generated the best average AUROC value.

All six single ML techniques are boosted by the AdaBoost algorithm. However, the boosting process might perform poorly if single classifiers are too complex for the available training data (Witten et al. 2016). Experiments in this research showed that the performance is improved only in the Naïve Bayes and MLP, which are the weakest performing single techniques, and they are still being outperformed by the remaining single techniques. The AdaBoost results of the remaining algorithms are not given as they do not improve the single versions.

Discussion of Findings

The literature review revealed 25 attributes that prominently impact occurrence of disputes and these attributes can be grouped under four categories as attributes related to project characteristics (PC), skills of the parties involved (S), changes (C), and delays (D). These 25 attributes were used to establish a conceptual model and this model was used to collect data. In a similar vein, the proposed conceptual model for predicting the occurrence of disputes can be used with different datasets to conduct similar research.

Chi-square tests showed that 14 of these attributes have significant association with dispute occurrence. Among the PC attributes, project location (PC1), value (PC2), and planned duration (PC3) were the selected attributes for the final prediction model. In a study that reviewed the literature to identify dispute factors, these three attributes were also reported among the prominent factors (Ilter 2012). On the other hand, construction type (PC4), contractor type (PC5), employer type (PC6), and contract type (PC7) were not significantly associated with dispute occurrence. Similarly, payment method (PC8), project delivery system (PC9), and level of design and construction complexity (PC10, PC11) attributes were also eliminated. This finding contradicts with the studies (such as Cheng et al. 2009; Chou and Lin 2013) that relate disputes with certain types of construction, project delivery system, etc.

Among the S attributes, relationship between parties / individuals (S1), previous experience with each other / reputation (S2), communication between parties (S4), working culture and skills of the represented party (S5-1) and the counter party (S5-2), response rate and communication skills of the counter party (S6-2), experience of the counter party (S7-2), project management and coordination skills of the represented party (S8-1), and the counter party (S8-2) were found to be significantly associated with dispute occurrence. It was interesting that the adequacy of dispute avoidance incentives (S3) was not found significant and left out of the prediction model. Although it was expected that organizational goals or reward mechanisms would increase individuals' motivation to avoid disputes (Diekmann and Girard 1995; Molenaar et al. 2000), findings in this study did not support this argument. This irregularity may be associated with the rooted adversarial relationships in the construction industry and even adequate avoidance incentives cannot prevent occurrence of disputes. Another interesting finding was related to response rate and communication skills of the represented party (S6-1). Although S4 and S6-2 attributes were significantly associated with dispute occurrence, it was found that S6-1 is not significantly associated. When responses to S6-1 attribute were reviewed, it is observed that the experts tend to overestimate their skills related to this attribute. Experts rated their response rate and communication skills for their organization (S6-1) as strong (Level 4) or very strong (Level 5) in 78.7% of the cases while this rate is 40.7% for the counter parties (S6-2).

The highest strength of association (Cramer's V = 0.576) is obtained for the changes (C1) attribute. Therefore, it is the most influential factor on dispute occurrence. This was also verified by the information gain ratio measure as C1 attribute generated the highest gain ratio (0.338). This finding is in accordance with the previous research stating that variations comprise one of the major causes of disputes (Kilian and Gibson 2005; Chen and Hsu 2007). According to their influence on dispute occurrence, the other influential attributes were S1 (Somers' d = -0.406; second highest gain ratio = 0.136), S4 (Somers' d = -0.370; third highest gain ratio = 0.115), S5-2 (Somers' d = -0.303; fourth highest gain ratio = 0.082), and S8-2 (Somers' d = -0.321; fifth highest gain ratio = 0.082). Therefore, majority of influential attributes were related to the parties involved. Considering dispute management is a process dominated by human factors (Cheung and Suen 2002), these results highlight the primary areas to focus in individual and organizational level. The negative signs in Somers' d values indicate that when the level of the attribute increases, the dispute rate decreases. Therefore, efforts directed to improve these areas will be beneficial in dispute avoidance. Other than the S attributes, the most influential project characteristics related attribute on dispute occurrence is the planned project duration (PC3) according to the gain ratio values. It is followed by another duration related attribute, delays (D1), which represents the ratio of extensions to planned duration.

In summary, PC1, PC2, PC3, S1, S2, S4, S5-1, S5-2, S6-2, S7-2, S8-1, S8-2, C1, and D1 were used in the prediction model. 10-fold CV results with 10 repeats showed that RBF kernel SVM, polynomial kernel SVM, and C4.5 algorithms are the three-best performing single techniques for the reviewed dataset with average prediction accuracies of 90.46%, 89.91%, and 88.98%, respectively. Hence, the SVM algorithm outperformed the remaining single techniques tested in this paper. The superiority of the SVM algorithm is not surprising considering that the algorithm was developed specifically for binary classification problems (Witten et al. 2016).

In pursue of an enhancement to classification performances, ensemble classifiers were developed. Experiments revealed that the best classifier for predicting the occurrence of disputes, which was obtained from the majority voting technique by combining classification decisions of the top three single techniques, can achieve 91.11% average prediction accuracy. Moreover, this classifier generated the best performance in Kappa, precision (the accuracy in prediction of disputed cases), and specificity (identifying undisputed cases correctly) measures. Although the highest AUROC value is obtained from the AdaBoost technique that boosted the Naïve Bayes classifier (0.954), the majority voting technique has the advantage of producing higher average accuracy compared with the boosted Naive Bayes classifier (88.06%), and thus, it is more desirable. In addition, the highest TP rate is obtained from the stacked classifier combining RBF and polynomial kernel SVM classifiers (0.931). However, as explained earlier, the TP rate of majority voting is only slightly lower (0.926), while it has better performance in

remaining evaluation criteria compared with the stacked classifier. Therefore, the final prediction model for predicting the occurrence of disputes in this paper is the classifier obtained from the majority voting technique.

Among limited empirical research on dispute occurrence prediction, Chou and Lin (2013) achieved a 10-fold CV prediction accuracy of 84.33% from the ensemble model that combined SVM, ANN, and C5.0 classifiers. They also achieved 85.60% precision, 95.26% sensitivity, 48.82% specificity, and 0.7229 AUROC values. The prediction model proposed in this study not only outperformed this classifier in terms of accuracy, but also it was capable of generating higher precision (93.70%), sensitivity (92.60%), specificity (88.40%), and AUROC (0.9050) values; although it should be noted that compared studies used different datasets and attributes. Chou et al. (2014) developed an SVM model on the same dataset to achieve 89.30% prediction success along with 94.67% precision, 74.24% sensitivity, 93.64% specificity, and 0.8364 AUROC values. This classifier was also outperformed by the proposed study. Although precision and specificity values seem higher in Chou et al. (2014), they were obtained from the best classifier with only a single trial and the variance in ML algorithms was not considered. On the other hand, the proposed study presents average results obtained from repeating each test 10 times. In another trial by repeating classification tests 10 times, Chou et al. (2016) achieved an average 10-fold CV accuracy of 83.92% using C5.0 algorithm. In the light of these comparisons, the proposed prediction model in this research achieved better performance in predicting dispute occurrence and it can firmly be concluded that results are promising.

The efforts to identify attributes that impact dispute occurrence showed that the literature relates several subjectively assessed attributes with the occurrence of disputes. The empirical dataset collected for this research is also based on the subjective judgments of the domain experts. For example, level of design complexity was subjectively assessed by the experts. This is regarded as the main limitation of the research. Another limitation is related to the data scarcity. Although the collected data is claimed to be quite representative, the number of projects is still limited due to access to such information, research duration, and budget. It can be increased to improve the generalization capability of the presented model. This data scarcity problem has also been highlighted by Yu (2007) by stating that historical data are scarce in nature for construction industry. Another limitation is the extent of the experimented techniques. Although performances of various ML techniques were compared, there are considerable classification techniques that are not evaluated in this research such as Random Forest, Convolutional Neural Networks, etc. with the potential to achieve an advancement in classification performance.

Conclusions

This study experimentally compared the performances of several single and ensemble ML techniques to predict the occurrence of disputes. In terms of average prediction accuracy, RBF kernel SVM (90.46%), polynomial kernel

SVM (89.91%), and C4.5 (88.98%) classifiers gave the best classification performances. However, two ensemble classifiers outperformed the single techniques by achieving 91.11% average accuracy, which are the stacked generalization technique that combined RBF and polynomial kernel SVMs, and the majority voting that combined all three-best performing single techniques. The superiority of the classifier obtained from majority voting technique in Kappa (0.806), precision (0.937), and specificity (0.884) measures made it become the best classifier in predicting the occurrence of disputes in this research.

The outcomes of this research can be valuable for professionals as it will avail early planning for taking necessary precautions, which may help reducing the effort, time, and cost of dispute management actions considerably. In addition, the research contributes to disputes literature with an empirical study that consider variations in project and organizational characteristics. Moreover, researchers can benefit from the conceptual model that constitutes a generic approach for dispute occurrence by using it to conduct similar research. As another contribution, the most influential factors on dispute occurrence are highlighted and results showed that occurrence of variations and unexpected events are the primary respondent of dispute occurrence. Moreover, the relationship between parties, communication between parties, and project management and coordination skills of the counter party have the highest impact on dispute occurrence, respectively. Researchers and construction professionals are advised to focus on enhancements on these areas to avoid dispute occurrence to a certain extent if not completely.

The next step of this research covers developing models for predicting the potential compensation and best resolution method in cases where disputes are inevitable. A potential tool combining the presented and forthcoming models can be beneficial by mitigating the detrimental effects of disputes. In addition, to mitigate the data scarcity problem in this research, researchers that will conduct similar studies may resort to the soft computing approach proposed by Yu (2007), which integrates fuzzy logic, learning ability of classifiers, and messy Genetic Algorithm.

Data Availability Statement

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request. Some or all data, models, or code generated or used during the study are proprietary or confidential in nature and may only be provided with restrictions. Company names, project names, and participant contact information cannot be provided to preserve anonymity of participants and to comply with legal issues such as privacy laws; instead, generic identification numbers have been assigned to each case.

References

Agresti, A. 2007. *An introduction to categorical data analysis*. New York: John Wiley & Sons.

595 Akoglu, H. 2018. “User’s guide to correlation coefficients.” *Turk. J. Emerg. Med.* 18 (3): 91–93.
596 https://doi.org/10.1016/j.tjem.2018.08.011.
597 Al-Khoder, A., and H. Harmouch. 2015. “Evaluating four of the most popular open sources and free data mining
598 tools.” *Int. J. Acad. Sci. Res.* 3 (1): 13–23.
599 Alpaydin, E. 2010. *Introduction to machine learning*. Cambridge, MA: MIT Press.
600 American Arbitration Association (AAA). 2018. “2017 annual report.” Accessed April 25, 2020.
601 http://www.adr.org/sites/default/files/document_repository/AAA_AnnualReport_Financials_2018.pdf.
602 An, S., U. Park, K. Kang, M.-Y. Cho, and H. Cho. 2007. “Application of support vector machines in assessing
603 conceptual cost estimates.” *J. Comput. Civ. Eng.* 21 (4): 259–264. https://doi.org/10.1061/(ASCE)0887-
604 3801(2007)21:4(259).
605 Arasu, B. S., B. J. B. Seelan, and N. Thamaraiselvan. 2020. “A machine learning-based approach to enhancing
606 social media marketing.” *Comput. Electr. Eng.* 86 (2020): 106723.
607 https://doi.org/10.1016/j.compeleceng.2020.106723.
608 Arditi, D., and T. Pulket. 2010. “Predicting the outcome of construction litigation using an integrated artificial
609 intelligence model.” *J. Comput. Civ. Eng.* 24 (1): 73–80. https://doi.org/10.1061/(ASCE)0887-
610 3801(2010)24:1(73).
611 Awwad, R., B. Barakat, and C. Menassa. 2016. “Understanding dispute resolution in the Middle East region from
612 perspectives of different stakeholder.” *J. Manage. Eng.* 32(6): 05016019.
613 https://doi.org/10.1061/(ASCE)ME.1943-5479.0000465.
614 Ayhan, M. 2019. Development of Dispute Prediction and Resolution Method Selection Models for Construction
615 Disputes. Doctoral Dissertation, Ankara, Turkey: Middle East Technical Univ.
616 Bilal, M., L. O. Oyedele, J. Qadir, K. Munir, S. O. Ajayi, O. O. Akinade, H. A. Owolabi, H. A. Alaka, and M.
617 Pasha. 2016. “Big data in the construction industry: a review of present status, opportunities, and future
618 trends.” *Adv. Eng. Inf.* 30 (2016): 500–521. https://doi.org/10.1016/j.aei.2016.07.001.
619 Chen, J. H., and S. C. Hsu. 2007. “Hybrid ANN-CBR model for disputed change orders in construction projects.”
620 *Autom. Constr.* 17 (1): 56–64. https://doi.org/10.1016/j.autcon.2007.03.003.
621 Cheng, M. Y., and Y. W. Wu. 2009. “Evolutionary support vector machine inference system for construction
622 management.” *Autom. Constr.* 18 (5): 597–604. https://doi.org/10.1016/j.autcon.2008.12.002.
623 Cheng, M. Y., H. C. Tsai, and Y. H. Chiu. 2009. “Fuzzy case-based reasoning for coping with construction
624 disputes.” *Expert Syst. Appl.* 36 (2): 4106–4113. https://doi.org/10.1016/j.eswa.2008.03.025.

Cheung, S. O., and H. C. H. Suen. 2002. “A multi-attribute utility model for dispute resolution strategy selection.” *Constr. Manage. Econ.* 20 (7): 557–568. https://doi.org/10.1080/01446190210157568.

Cheung, S. O., R. F. Au-Yeung, and V. W. K. Wong. 2004. “A CBR based dispute resolution process selection system.” *Int. J. IT Archit., Eng. Constr.* 2 (2): 129–145.

Cheung, S. O., and K. H. Y. Pang. 2013. “Anatomy of construction disputes.” *J. Constr. Eng. Manage.* 139 (1): 15–23. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000532.

Chou, J. S. 2012. “Comparison of multilabel classification models to forecast project dispute resolutions.” *Expert Syst. Appl.* 39 (11): 10202–10211. https://doi.org/10.1016/j.eswa.2012.02.103.

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.

Chou, J. S., C. Tsai, and Y. Lu. 2013a. “Project dispute prediction by hybrid machine learning techniques.” *J. Civ. Eng. Manage.* 19 (4): 505–517. https://doi.org/10.3846/13923730.2013.768544.

Chou, J. S., M. Y. Cheng, and Y. W. Wu. 2013b. “Improving classification accuracy of project dispute resolution using hybrid artificial intelligence and support vector machine models.” *Expert Syst. Appl.* 40 (6): 2263–2274. https://doi.org/10.1016/j.eswa.2012.10.036.

Chou, J. S., M. Y. Cheng, Y. W. Wu, and A. D. Pham. 2014. “Optimizing parameters of support vector machine using fast messy genetic algorithm for dispute classification.” *Expert Syst. Appl.* 41 (8): 3955–3964. https://doi.org/10.1016/j.eswa.2013.12.035.

Chou, J. S., S. C. Hsu, C. W. Lin, and Y. C. Chang. 2016. “Classifying influential information to discover rule sets for project disputes and possible resolutions.” *Int. J. Proj. Manage.* 34 (8): 1706–1716. https://doi.org/10.1016/j.ijproman.2016.10.001.

Dalton, D., and N. Shehadeh. 2003. “Statistical modelling of claims procedures and construction conflicts.” In P. Fenn and R. Gameson (Eds.), *Construction Conflict Management and Resolution*, 275–285. London: Routledge. https://doi.org/10.4324/9780203474396-27.

Diekmann, J. E., and M. J. Girard. 1995. “Are contract disputes predictable?” *J. Constr. Eng. Manage.* 121 (4): 355–363. https://doi.org/10.1061/(ASCE)0733-9364(1995)121:4(355).

Engineering News-Record (ENR). 2018. “ENR's top 250 international contractors.” Accessed April 25, 2020. https://www.enr.com/toplists/2018-Top-250-International-Contractors-1.

Febriantono, M. A., S. H. Pramono, R. Rahmadwati, and G. Naghdy. 2020. “Classification of multiclass imbalanced data using cost-sensitive decision tree C5.0.” *IAES Int. J. Artif. Intel.* 9 (1): 65–72. https://doi.org/10.11591/ijai.v9.i1.pp65-72.

Fenn, P. 2007. “Predicting construction disputes: An aetiological approach.” In Vol. 160(MP2) of *Proc., the ICE - Management, Procurement and Law*, 69–73. U.K.: ICE Publishing.

Frank, E., H. A. Mark, and I. H. Witten. 2016. “The WEKA Workbench.” *Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"* 4th ed., Burlington, MA: Morgan Kaufmann.

Hong Kong International Arbitration Center (HKIAC). 2018. “HKIAC annual report 2018 reflections.” Accessed April 25, 2020. https://www.hkiac.org/sites/default/files/annual_report/annual%20report%203463-7390-6190%20v.4.pdf.

Hssina, B., A. Merbouha, H. Ezzikouri, and M. Erritali. 2014. “A comparative study of decision tree ID3 and C4.5.” *Int. J. Adv. Comp. Sci. Appl.* 4 (2): 13–19.

Ilter, D., and A. Dikbas. 2009. “An investigation of the factors influencing dispute frequency in construction projects.” In *Proc., RICS Int. Res. Conf. (COBRA 2009)*, 1496–1504. West Yorkshire: Emerald Publishing.

Ilter, D. 2012. “Identification of the relations between dispute factors and dispute categories in construction projects.” *Int. J. Law in the Built Environment* 4 (1): 45–59. https://doi.org/10.1108/17561451211211732.

Kilian, J. J., and G. E. Gibson. 2005. “Construction litigation for the U.S. naval facilities engineering command, 1982–2002.” *J. Constr. Eng. Manage.* 131 (9): 945–952. https://doi.org/10.1061/(ASCE)0733-9364(2005)131:9(945).

Liao, S. H., P. H. Chu, and P. Y. Hsiao. 2012. “Data mining techniques and applications - a decade review from 2000 to 2011.” *Expert Syst. Appl.* 39 (12): 11303–11311. https://doi.org/10.1016/j.eswa.2012.02.063.

Love, P. E. D., P. R. Davis, J. Ellis, and S. O. Cheung. 2010. “Dispute causation: Identification of pathogenic influences in construction.” *Eng. Constr. Archit. Manage.* 17 (4): 404–423. https://doi.org/10.1108/09699981011056592.

Mahfouz, T., A. Kandil, and S. Davlyatov. 2018. “Identification of latent knowledge in differing site condition (DSC) litigations.” *Autom. Constr.* 94: 104–111. https://doi.org/10.1016/j.autcon.2018.06.011.

McHugh, M. L. 2012. “Interrater reliability: The Kappa statistic.” *Biochemia Medica*, 22(3), 276–282.

McHugh, M. L. 2013. “The chi-square test of independence.” *Biochemia Medica*, 23(2), 143–149. https://doi.org/10.11613/BM.2013.018.

683 Mokhtar, S. H. M., and S. A. Rahman. 2017. “The roles of big data and knowledge management in business
684 decision making process.” *Int. J. Acad. Res. Bus. Soc. Sci.* 7 (12): 422–428.
685 https://doi.org/10.6007/IJARBSS/v7-i12/3623.
686 Molenaar, K., S. Washington, and J. Diekmann. 2000. “Structural equation model of construction contract dispute
687 potential.” *J. Constr. Eng. Manage.* 126 (4): 268–277. https://doi.org/10.1061/(ASCE)0733-
688 9364(2000)126:4(268).
689 Parikh, D., G. J. Joshi, and D. A. Patel. 2019. “Development of prediction models for claim cause analyses in
690 highway projects.” *J. Leg. Aff. Dispute Resolut. Eng. Constr.* 11(4): 04519018.
691 https://doi.org/10.1061/(ASCE)LA.1943-4170.0000303.
692 Pollock III, P. H. 2011. *An SPSS companion to political analysis.* Washington DC: CQ Press.
693 Pulket, T., and D, Arditi. 2009. “Construction litigation prediction system using ant colony optimization.” *Constr.*
694 *Manage. Econ.* 27 (3): 241–251. https://doi.org/10.1080/01446190802714781.
695 Soni, S., M. Pandey, and S. Agrawal. 2017. “Conflicts and disputes in construction projects: an overview.” *Int. J.*
696 *Eng. Res. Appl.* 7 (06): 40–42. https://doi.org/10.9790/9622-0706074042.
697 Sonmez, R., and B. Sozgen. 2017. “A support vector machine method for bid/no bid decision making.” *J. Civ.*
698 *Eng. Manage.* 23 (5): 641–649. https://doi.org/10.3846/13923730.2017.1281836.
699 Ustuner, Y. A., and E. Tas. 2019. “An examination of the mediation processes of international ADR institutions
700 and evaluation of the Turkish construction professionals’ perspectives on mediation.” *Eurasian J. Soc. Sci.* 7
701 (4): 11–27. https://doi.org/10.15604/ejss.2019.07.04.002.
702 Vanwinckelen, G., and H. Blockeel. 2012. “On estimating model accuracy with repeated cross-validation.” In
703 *Proc., 21st Belgian-Dutch Conf. on ML,* 39–44. Ghent, Belgium: Benelearn 2012 Organization Committee.
704 Weisburd, D., and C. Britt. 2007. “Chapter 13: Measures of association for nominal and ordinal variables”. In:
705 *Statistics in Criminal Justice*, 335–380. Boston, MA: Springer.
706 Witten, H. W., E. Frank, M. A. Hall, and C. J. Pal. 2016. *Data mining: Practical machine learning tools and*
707 *techniques*. Burlington, MA: Morgan Kaufmann.
708 Yousefi, V., S. H. Yakhchali, M. Khanzadi, E. Mehrabanfar, and J. Šaparauskas. 2016. “Proposing a neural
709 network model to predict time and cost claims in construction projects.” *J. Civ. Eng. Manage.* 22 (7): 967–
710 978. https://doi.org/10.3846/13923730.2016.1205510.
711 Yu, W. D. 2007. “Hybrid soft computing approach for mining of complex construction databases.” *J. Comput.*
712 *Civ. Eng.* 21 (5): 343–352. https://doi.org/10.1061/(ASCE)0887-3801(2007)21:5(343).

Table 1. Project Characteristics (PC) Attributes – Categorical Labels and Frequencies

ID	Attribute	Categorical Label		Frequency	Relative Frequency (%)
PC1	Project Location	1	Domestic	80	74.1
		2	International	28	25.9
PC2	Project (Contract) Value	1	< 10 million $	44	40.7
		2	10-100 million $	35	32.4
		3	> 100 million $	29	26.9
PC3	Planned Project Duration	1	< 1 year	31	28.7
		2	1 – 2 years	37	34.3
		3	2 – 3years	21	19.4
		4	> 3 years	19	17.6
PC4	Type of Construction	1	Housing	18	16.7
		2	Commercial	10	9.3
		3	Industrial	12	11.1
		4	Transportation	17	15.7
		5	Power Plants & Lines	8	7.4
		6	Water Supplies & Reservoirs	10	9.3
		7	Sports & Cultural & Educational	11	10.2
		8	Medical	7	6.5
		9	Public	6	5.6
		10	Soil Works	9	8.3
PC5	Type of Contractor	1	Single	88	81.5
		2	Joint Venture	11	10.2
		3	Consortium	9	8.3
PC6	Type of Employer	1	Public	52	48.1
		2	Private	43	39.8
		3	Public-Private-Partnership	13	12.0
PC7	Type of Contract	1	Private Contracts	53	49.1
		2	Public Procurement	36	33.3
		3	FIDIC Red	10	9.3
		4	FIDIC Silver & Yellow	9	8.3
PC8	Payment Method	1	Fixed (Lump-Sum)	58	53.7
		2	Unit Price	50	46.3
PC9	Project Delivery System	1	Design-Bid-Build	67	62.0
		2	Design-Build	26	24.1
		3	Engineering-Procurement-Construction	15	13.9
PC10	Level of Design Complexity	1	Very Low	13	12.0
		2	Low	16	14.8
		3	Moderate	20	18.5
		4	High	37	34.3
		5	Very High	22	20.4
PC11	Level of Construction Complexity	1	Very Low	9	8.3
		2	Low	15	13.9
		3	Moderate	19	17.6
		4	High	38	35.2
		5	Very High	27	25.0

Table 2. Skills (S) Attributes – Levels and Frequencies

ID	Attribute	Levels	Frequency	Relative Frequency (%)	ID	Attribute	Levels	Frequency	Relative Frequency (%)
S1	Relationship between parties / individuals	Level 1	10	9.3	S6-1	Response rate & communication skills of the represented party	Level 1	10	9.3
		Level 2	14	13.0			Level 2	14	13.0
		Level 3	12	11.1			Level 3	12	11.1
		Level 4	48	44.4			Level 4	48	44.4
		Level 5	24	22.2			Level 5	24	22.2
S2	Previous experience with each other / Reputation	Level 1	2	1.9	S6-2	Response rate & communication skills of the counter party	Level 1	19	17.6
		Level 2	7	6.5			Level 2	21	19.4
		Level 3	20	18.5			Level 3	24	22.2
		Level 4	42	38.9			Level 4	26	24.1
		Level 5	37	34.3			Level 5	18	16.7
S3	Dispute avoidance incentive	Level 1	46	42.6	S7-1	Experience of the represented party	Level 1	1	0.9
		Level 2	3	2.8			Level 2	3	2.8
		Level 3	16	14.8			Level 3	15	13.9
		Level 4	21	19.4			Level 4	33	30.6
		Level 5	22	20.4			Level 5	56	51.9
S4	Communication between parties	Level 1	7	6.5	S7-2	Experience of the counter party	Level 1	11	10.2
		Level 2	18	16.7			Level 2	16	14.8
		Level 3	25	23.1			Level 3	22	20.4
		Level 4	34	31.5			Level 4	30	27.8
		Level 5	24	22.2			Level 5	29	26.2
S5-1	Working culture & skills of the represented party	Level 1	1	0.9	S8-1	Project management & coordination skills of the represented party	Level 1	1	0.9
		Level 2	7	6.5			Level 2	4	3.7
		Level 3	20	18.5			Level 3	22	20.4
		Level 4	45	41.7			Level 4	50	46.3
		Level 5	35	32.4			Level 5	31	28.7
S5-2	Working culture & skills of the counter party	Level 1	18	16.7	S8-2	Project management & coordination skills of the counter party	Level 1	10	9.3
		Level 2	17	15.7			Level 2	26	24.1
		Level 3	29	26.9			Level 3	32	29.6
		Level 4	27	25.0			Level 4	31	28.7
		Level 5	17	15.7			Level 5	9	8.3

Table 3. Changes (C) Attribute – Categorical Labels and Frequencies

ID	Attribute	Categorical Label		Frequency	Relative Frequency (%)
C1	Changes	0	No	67	62.0
		1	Yes	41	38.0

Table 4. Delays (D) Attribute – Categorical Labels and Frequencies

ID	Attribute	Categorical Label		Frequency	Relative Frequency (%)
D1	Delays	1	Ratio = 0%	44	40.7
		2	Ratio 0-20%	24	22.2
		3	Ratio 20-40%	17	15.7
		4	Ratio > 40%	23	21.3

Table 5. The Chi-Square Test Results of Project Characteristics (PC) Attributes

Attribute	Categories	p-value	Dispute Occurrence Rate (%)	Selected for Final Model	Strength of Association
PC1 – Project Location		0.037		YES	Cramer's V
	Domestic		58.8		0.215
	International		82.1		
PC2 – Project (Contract) Value		0.003		YES	Somers' d
	< 10 million $		47.7		0.259
	10-100 million $		68.6		
	> 100 million $		86.2		
PC3 – Planned Project Duration		0.000		YES	Somers' d
	< 1 year		41.9		0.286
	1-2 years		62.2		
	2-3 years		71.4		
	> 3 years		100.0		
PC4 – Type of Construction		0.157		NO	-
PC5 – Type of Contractor		0.749		NO	-
PC6 – Type of Employer		0.961		NO	-
PC7 – Type of Contract		0.074		NO	-
PC8 – Payment Method		0.842		NO	-
PC9 – Project Delivery System		0.957		NO	-
PC10 – Level of Design Complexity		0.938		NO	-
PC11 – Level of Construction Complexity		1.000		NO	-

Table 6. The Chi-Square Test Results of Skills (S) Attributes

Attribute	p-value	Selected for Final Model	Strength of Association
S1 – Relationship between parties / individuals	0.000	YES	Somers' d -0.406
S2 – Previous Experience with each other / Reputation	0.007	YES	Somers' d -0.185
S3 – Dispute avoidance incentive	0.158	NO	-
S4 – Communication between parties	0.000	YES	Somers' d -0.370
S5-1 – Working culture & skills of the represented party	0.012	YES	Somers' d -0.162
S5-2 – Working culture & skills of the counter party	0.000	YES	Somers' d -0.303
S6-1 – Response rate & communication skills of the represented party	0.228	NO	-
S6-2 – Response rate & communication skills of the counter party	0.000	YES	Somers' d -0.280
S7-1 – Experience of the represented party	0.085	NO	-
S7-2 – Experience of the counter party	0.001	YES	Somers' d -0.233
S8-1 – Project management & coordination skills of the represented party	0.006	YES	Somers' d -0.199
S8-2 – Project management & coordination skills of the counter party	0.000	YES	Somers' d -0.321

Table 7. The Chi-Square Test Results of Changes (C) Attribute

Attribute	Categories	p-value	Dispute Occurrence Rate (%)	Selected for Final Model	Strength of Association
C1 – Changes		0.000		YES	Cramer's V
	Yes		100.0		0.576
	No		43.3		

Table 8. The Chi-Square Test Results of Delays (D) Attribute

Attribute	Categories	p-value	Dispute Occurrence Rate (%)	Selected for Final Model	Strength of Association
D1 – Delays		0.002		YES	Somers' d
	Ratio = 0%		47.7		0.232
	Ratio 0-20%		62.5		
	Ratio 20-40%		94.1		
	Ratio > 40%		78.3		

Table 9. Gain Ratio Values of the Input Variables with respect to Dispute Occurrence

Rank	Attribute ID	Gain Ratio Value	Selected for Final Model Based on Chi-Square Results	Rank	Attribute ID	Gain Ratio Value	Selected for Final Model Based on Chi-Square Results
1	C1	0.338	YES	15	S3	0.030	NO
2	S1	0.136	YES	16	PC7	0.030	NO
3	S4	0.115	YES	17	PC4	0.027	NO
4	S5-2	0.082	YES	18	S7-1	0.018	NO
5	S8-2	0.082	YES	19	PC11	0.015	NO
6	PC3	0.081	YES	20	S6-1	0.015	NO
7	S6-2	0.079	YES	21	PC10	0.007	NO
8	S5-1	0.068	YES	22	PC5	0.005	NO
9	D1	0.055	YES	23	PC9	0.001	NO
10	S7-2	0.055	YES	24	PC6	0.001	NO
11	PC2	0.053	YES	25	PC8	0.000	NO
12	S8-1	0.046	YES				
13	PC1	0.044	YES				
14	S2	0.039	YES				

Table 10. Parameter Configurations for the Utilized ML techniques

Algorithm	Parameter	Search Range
Naïve Bayes	No parameter optimization is required	
kNN	k value	1-100 (in step size of 1)
	Distance measurement function	Chebyshev, Euclidean, Manhattan, Minkowski
	Distance weighting method	Equal, Inverse (1/weight), Similarity (1-weight)
C4.5	Pruning	True, False
	Reduced error pruning	True, False
	Subtree raising	True, False
	Confidence threshold factor for pruning	[0.01-0.50] (in step size of 0.01)
	Minimum number of instances per leaf	[1-10] (in step size of 1)
	Number of folds for pruning	[2-5] (in step size of 1)
	Use Laplace for counts at leaves	True, False
MLP	Number of hidden layers	0, 1, 2, (total number of inputs and outputs) / 2
	Number of epochs (cycles)	500, 1000
	Momentum	[0.1-0.9] (in step size of 0.1)
	Learning rate	[0.1-0.9] (in step size of 0.1)
Polynomial SVM	Penalty parameter C	$[2^{-2}$-$2^{15}]$ (in exponentially growing sequence)
	Exponent	[1-10] (in step size of 1)
RBF kernel SVM	Penalty parameter C	$[2^{-2}$-$2^{15}]$ (in exponentially growing sequence)
	RBF kernel gamma	$[2^{-15}$-$2^{4}]$ (in exponentially growing sequence)
Voting Technique	Combination rule	Majority voting, Average of Probabilities Voting
Stacked Generalization	Base learner	Top 3 performing single classifiers
	Meta-learner	Remaining single techniques (base-learner and meta-learner cannot be the same technique)
AdaBoost	Number of iterations to be performed	10
	Boosting mechanism	True (use resampling), False (use reweighting)

Table 11. Confusion Matrix for Binary Classification

Class	Predicted Class: Disputed	Predicted Class: Undisputed
Actual Class: Disputed	TP	FN
Actual Class: Undisputed	FP	TN

Table 12. 10-Times 10-Fold CV Performance of Single Classifiers

Algorithm	Avg. Accuracy (%)	%95 CI Accuracy (%)	Avg. Kappa	Avg. Precision	Avg. Recall (TP Rate)	Avg. Specificity	Avg. AUROC
Naive Bayes	87.50	[86.60-88.40]	0.728	0.912	0.893	0.843	**0.953**
kNN	87.69	[86.65-88.72]	0.737	**0.931**	0.874	**0.881**	0.928
C4.5	88.98	[87.26-90.70]	0.761	0.927	0.901	0.868	0.947
MLP	83.52	[82.06-84.98]	0.641	0.879	0.866	0.779	0.894
Poly. Kernel SVM	89.91	[88.85-90.96]	0.777	0.917	**0.929**	0.845	0.887
RBF kernel SVM	**90.46**	[89.17-91.75]	**0.790**	0.925	**0.929**	0.861	0.894

Note: Bold values show the best performance obtained in the corresponding measure

Table 13. 10-Times 10-Fold CV Performance of Ensemble Classifiers

Algorithm	Avg. Accuracy (%)	%95 CI Accuracy (%)	Avg. Kappa	Avg. Precision	Avg. Recall (TP Rate)	Avg. Specificity	Avg. AUROC
Majority Voting	**91.11**	[89.93-92.29]	**0.806**	**0.937**	0.926	**0.884**	0.905
Stacking: RBF kernel SVM + Poly. kernel SVM	**91.11**	[89.78-92.44]	0.805	0.932	**0.931**	0.874	0.903
AdaBoost: Naïve Bayes	88.06	[87.20-88.91]	0.738	0.908	0.907	0.832	**0.954**
AdaBoost: MLP	83.70	[81.68-85.73]	0.651	0.895	0.849	0.816	0.908

Note: Bold values show the best performance obtained in the corresponding measure