Received: 8 September 2022 | Accepted: 7 October 2023

DOI: 10.1111/mice.13114

INDUSTRIAL APPLICATION

COMPUTER-AIDED CIVIL AND INFRASTRUCTURE ENGINEERING WILEY

An automated machine learning approach for classifying infrastructure cost data

Daniel Adanza Dopazo[1] | **Lamine Mahdjoubi**[1] | **Bill Gething**[1] | **Abdul-Majeed Mahamadu**[2]

[1]School of architectue and environment, University of the West England, Bristol, England, UK

[2]The Bartlett School of Sustainable Construction, University College London, London

Correspondence
Daniel Adanza Dopazo, University of the West England, Bristol, School of architecture and environment, Coldharbour Ln, Stoke Gifford, Bristol BS16 1QY, UK.
Email: Daniel.Dopazo@uwe.ac.uk

Funding information
Innovate UK, Grant/Award Number: 45382; Transport Infrastructure Efficiency Strategy Living Lab, Grant/Award Number: 45382

[Correction added on November 02, 2023, after first publication: The affiliation of fourth author has been changed.]

Abstract

Data on infrastructure project costs are often unstructured and lack consistency. To enable costs to be compared within and between organizations, large amounts of data must be classified to a common standard, typically a manual process. This is time-consuming, error-prone, inconsistent, and subjective, as it is based on human judgment. This paper describes a novel approach for automating the process by harnessing natural language processing identifying the relevant keywords in the text descriptions and implementing machine learning classifiers to emulate the expert's knowledge. The task was to identify "extra over" cost items, conversion factors, and to recognize the correct work breakdown structure (WBS) category. The results show that 94% of the "extra over" cases were correctly classified, and 90% of cases that needed conversion, correctly predicting an associated conversion factor with 87% accuracy. Finally, the WBS categories were identified with 72% accuracy. The approach has the potential to provide a step change in the speed and accuracy of structuring and classifying infrastructure cost data for benchmarking.

1 | INTRODUCTION

The increasingly bigger amount of data in infrastructure project cost, its inconsistencies in structure, and the variety of formats, even within a single organization, makes the comparison process, the analysis, and the decision-making tasks difficult and unreliable.

The problem does not have an easy solution, and extracting and reclassifying this information into a common standard involves the manipulation of large amounts of information systematically (Ahiaga-Dagbui & Smith, 2012). This task is still largely performed manually by many construction companies. Due to the size of the data involved, this task is time-consuming, error-prone, and may also be inconsistent and subjective since it is based on human judgment (Martínez-Rojas et al., 2015).

Additionally, processing large amounts of often disparate cost data present additional challenges such as: including varying formats and unstructured content of cost documents and lack of standardization of the data schemas, as well as poor cost classification practices across the industry (Matthews et al., 2022; NIC, 2021).

This, in turn, results in severe challenges in data analytics and benchmarking, within and between organizations let alone across the wider infrastructure industry. In fact, according to the Data Management Association, organizations spend between 10% and 30% of their revenue on managing data quality issues (GDQH, 2021). This includes costs associated with pre-processing and the general loss of information due to inconsistency and variable quality.

A recent report confirmed that 32% of construction project data is often of poor quality, being inaccurate, unusable, or inconsistent (Thomas & Bowman, 2021). According to the same report, only 30% of organizations implement systems for improving data quality as well as usability. Among other factors, this is often due to reliance on manual data pre-processing, which is typically cumbersome, inefficient, slow, and prone to errors.

Benchmarking and cost analysis in the infrastructure construction sector continue to be fraught with data inconsistency issues thus preventing the sector from unlocking the benefits of cost intelligence (NIC, 2021).

The process of benchmarking transport infrastructure project costs is particularly essential for cost planning and estimation and thereby for driving efficiency and value in the delivery of projects (Elmousalami, 2020).

But once again, the lack of consistency in cost reporting practices across the infrastructure construction supply chain and client organizations presents a huge challenge for the mining of cost information in consistent formats and structures that enable robust comparison of like-for-like during benchmarking (Chen et al., 2019).

This emanates from differences in measurement methods, cost breakdown structures, and data collection techniques, as well as discrepancies in the granularity and classification of cost data (Martínez-Rojas et al., 2015).

Whereas some organizations in this field have approaches for agglomerating cost data, very often they need to deal with inconsistencies in data received from their supply chain. The processing of such data into a consistent format for further analysis can therefore be daunting due to time commitment and the need for expert professional judgment in the manual process of reclassifying such data (Matthews, 2022; NIC, 2021).

Cost information submitted by the infrastructure project supply chain often includes descriptions and categorization codes that are bespoke to the organizations submitting and potentially using cost reporting standards that are inconsistent with the cost breakdown structure adopted by the client organization for benchmarking or analysis.

This prevails despite efforts for cost reporting standardization with frameworks such as the International Cost Measurement Standard (ICMS, 2021). In practice, many organizations still use data structures specific to their profession, sector, country, organization, or even department within an organization.

As a result of these contextual differences, some cost breakdown structures are based on construction trades, and asset types, while some are based on construction elements or activities (Chen et al., 2019). Even where two organizations use the same "standards," there can still be discrepancies in the way individual projects or professionals present the data.

Thus, in order to make cost datasets more usable, it is important to reclassify cost items in an agreed-consistent way as well as easily identify items that require further processing to improve consistency and data quality.

Automation offers the potential significantly to streamline the speed and enhance the accuracy when extracting and analyzing infrastructure data (Matthews et al., 2022).

A number of machine learning (ML)-based techniques have been proposed for automating data mining in several fields, yet for the construction and infrastructure industries, this work has so far attracted little attention. (NIC, 2021; Wu et al., 2022).

To contribute to this field, this paper presents an automated method for the automatic classification of infrastructure maintenance cost data. First, it tokenizes the assets' descriptions at a word level. Second, it implements one-hoot encoding to identify relevant keywords to finally perform text classification, combining its results with three ML algorithms to support benchmarking and cost analysis in this sector.

The particular task tackled by this study was to (1) identify items where the units used for the original submission differed from the units used for the category to which the item was allocated in the agreed common standard (and if so, to identify any corresponding conversion factor), (2) to identify "extra over" items to isolate them in the aggregation of data during benchmarking, and (3) to automatically classify the work breakdown structure (WBS) category of each registered cost item. "Extra over" in construction costing refers to additional costs of items that have already been measured or presented in cost reporting documents such as bills of quantities (BOQs; Gorse et al., 2012). Whereas the costs are normally included in the analysis, the quantities would generally be excluded to avoid duplication since they share the same quantity as the parent cost item.

This paper reports on the development of a method with a novel approach, presenting a full-fledged system encompassing data wrangling, natural language processing (NLP) techniques implementing word tokenization and one-hoot encoding to extract relevant information out of the text descriptions, and the capabilities of three ML models to provide classifications. The solution has proved to obtain consistent and unbiased outputs when applied to a real-case scenario.

2 | RELATED WORK

In recent years, increasing attention has been paid to the usage of knowledge engineering and data mining for regression and classification in the construction sector. Many such studies have focused on the prediction and estimation of costs or cost performance (Elmousalami, 2020; Pham et al., 2021; Tayefeh Hashemi et al., 2020).

Other studies focus on feature engineering and case-based reasoning to infer the most important factors that influence the costs (Ji et al., 2019) or to assess the impact of these factors (Ahiaga-Dagbui & Smith, 2012). For that purpose, an artificial neural network has been implemented to build final cost estimation models.

However, despite the complexity and the novelty that brings the neural networks to the field, it has been demonstrated that these algorithms underperform more traditional ML algorithms in many scenarios (Bodendorf et al., 2021).

Alternatively, others implemented different regression algorithms. Within them, the most impactful and recent approaches will be highlighted:

First, Shoar et al. (2022) aim to establish a correlation between the risk factors and the ML models. For that purpose, a Bayesian neural network and a random forest are being compared.

Second, the novel approach by Florez-Perez et al. (2022) integrates the algorithms K-nearest neighbor, a deep neural network, logistic regression, and support vector machine (SVM) to discover the mapping between different subjective factors and the construction task productivity.

Third, the paper by Wang et al. (2022) presents an innovative approach to provide semantic classification, providing rich information for automatic construction, including remote construction and route planning of the mobile robotic systems.

Fourth, Akinosho et al. (2020) focus on assessing the main challenges and the results of many studies implementing deep learning in construction projects. One of the main inferences that can be extracted is the fact that the data pre-processing, data cleansing, and the hyper-parameter Turing processes seemed to be more impactful on the results than on the selection of the algorithms to provide predictions.

Finally, some other more former but impactful approaches are worth mentioning such as the implementation of a neurodynamic model for scheduling/cost optimization (Adeli & Karim, 1997) or its implementation for the optimization of the costs in composite beams (Adeli & Kim, 2001).

These studies have however primarily focused their attention on cost estimation and on case-based reasoning rather than the automation of the data processing or the information extraction overall.

A number of these studies have also focused on automating the classification of cost data. For example, Chen et al. (2019) proposed the combination that encompasses the accountant surveyor's knowledge with five different binary classifiers for automated cost analysis of priced BOQs data by classifying cost items into their trade-based categories.

Alternatively, Martínez-Rojas et al. (2015) developed a web-based tool for BOQ data classification using expert knowledge, text data extraction, and multi-criteria aggregation.

Many other approaches focused on solving non-cost-specific problems including the usage of semantic reasoning for detecting consistency in project scheduling (Zhao et al., 2020). Whereas other applications include automatic text categorization taking project reports as inputs (Sebastiani, 2002).

The benefits of combining NLP with classifiers for construction sector applications are being demonstrated in Wu et al. (2022), where 125 different research projects were reviewed applying a wide range of classifiers mainly grouped into four categories: syntactic parsing, heuristic rules, deep learning, and ML. Although it is important to be skeptical when comparing studies with different approaches, the paper demonstrates the increased interest in the field.

Otherwise, other approaches prefer the usage of semantic reasoning to extract construction methods from schedules with the main goal of ensuring consistency (Zhao et al., 2020) or to retrieve similar cases to seek efficiency in costs (Zou et al., 2017). The approach is based on combining two NLP techniques: vector space model and semantic query expansion. Finally, Rico-Juan, et al. (2019) similarly describe techniques for detecting inconsistencies between the text and graphical models.

Finally, quite aligned with the scope of this paper is Lin et al. (2016) where an approach based on NLP is presented for smart data retrieval in a scenario where big quantities of information are being presented for building information modeling.

Despite the proliferation of these studies, robust approaches for automating pre-processing tasks remain scarce, especially for the purposes of identifying inconsistencies in cost data and classifying instances thereby improving the quality of data mining.

The complexity of replicating the expert knowledge and the big data variability makes the pre-processing task a difficult solution to implement.

ML algorithms have great potential in this context. Their continuous learning capability makes them more accurate, especially in scenarios where large quantities of data need processing (Bottou, 2014). More specifically, decision trees offer significant benefits due to their transparency and robustness for text classification (de Ville, 2013). Similar text analysis approaches show significant power for representing expert knowledge through computer logic (Matthews et al., 2022; Soibelman et al., 2008).

More specifically, three ML algorithms were implemented in this solution for benchmarking and comparison: The first algorithm was the SVM:

The implementation of SVM in construction management projects has many possibilities. Examples include a decision-making system that is able to emulate knowledge experts' logic through the combination of SVM, genetic algorithms, and fuzzy logic (Cheng & Roy, 2010); an evolutionary SVM inference model to solve a wide range of complex construction management problems (Cheng & Wu, 2009); and the development of a hybrid risk-prediction framework to enhance the safety of the metro station construction (Liu et al., 2020).

The second chosen algorithm for classification was the random forest: They have been widely used in construction management for many different purposes, such as to predict construction and demolition waste generation (Cha et al., 2020), to develop a method that integrates the fractional calculus with random forest for the activity recognition of the construction equipment (Langroodi et al., 2021), or to compare the performance of the random forest with neural networks when predicting the building energy consumption (Ahmad et al., 2017), the study shows that both options are valid in that specific scenario with the neural networks performing marginally better.

Finally, the third chosen classifier was the stochastic gradient descendant (SGD): The algorithm has been used in different approaches when applied to the field, such as: developing a predicting model based on polynomial semantic indexing (Minoura et al., 2013); evaluating the conditioning factors of slope instability and the continuous changes in the generation of landslide inventory maps (Ramos-Bernal et al., 2021); or combining them with a genetic algorithm to propose a mathematical model of a laboratory-scale plant for slaughterhouse effluent bio digestion for biogas production (Martinez et al., 2012).

2.1 | The novelty of the method

In comparison with the state of the art, it can be inferred that the novelty of the presented paper relies on the following underpinnings:

(1) *The scope of the problem*: To our knowledge, there are no approaches presenting a method for automatically classifying manually included attributes and extracting relevant keywords based on text descriptions with the usage of already existing technologies.
(2) *The completeness of the solution*: From the reviewed literature, none of the studies presents a full-fledged methodology encompassing processes such as data wrangling, text analysis, and a reliable classification provided by three ML algorithms.
(3) *The emulation of experts'knowledge*: The emulation of the accountant surveyors' knowledge is a very unique and hence innovative task. To achieve this, a deep understanding of the classification costs combined with NLP techniques such as word tokenization and one-hoot encoding has been needed.
(4) *Its strong validation*: Although it is not an innovation in itself, it is important to highlight the validation of the method. From the analyzed literature, very few studies contain such an extent dataset composed of 59,140 real assets whose costs have been previously reviewed and classified by accountant surveyors. Additionally, a well-known client named "National Highways" has implemented the designed method that proves the efficacy of the approach implemented in a real-case scenario.
(5) *Superior accuracy than its peers*: Although it is important to be cautious when comparing studies implemented on different scenarios, it is also worth to stress that the presented approach overperforms its comparables, and this will be explained in detail in the Discussion section.

3 | MATERIALS AND METHODS

Road infrastructure organizations spend considerable amounts of time manually reclassifying project cost data, submitted by their supply chain in a variety of formats and structures, which although broadly based on the industry's Method of Measurement for Highway Works (MMHW, 2009), differ in detail from the client organization's standard cost reporting "WBS".

The paper suggests a method allowing for reclassification to a fully consistent standard in an efficient manner, harnessing the capabilities brought by NLP, automatizing the accountant surveyors' work, and supporting further analysis and benchmarking.

In this research, a large dataset of infrastructure work items, manually pre-classified by accountant surveyors depending on their cost categories and the features of the assets, was used to develop and train three ML models to replicate the experts' knowledge.

The input data essentially consisted of descriptions of work elements, as submitted by the supply chain, and a corresponding reclassification to the client's WBS. Actual costs and quantities and other project information were not included in the dataset in the interests of confidentiality.

In some cases, the reclassification introduced a further layer of complexity, in addition to the challenge of selecting the appropriate reclassification category, in that the unit of measure used for the original classification sometimes differs from the units used for the reclassified element.

For example, the original data may be expressed in square meters, whereas the reclassified data must be expressed in cubic meters. This means that the description of the work element must be interrogated using text mining and data science to identify syntax and information in the text to establish an appropriate conversion factor to be applied to the associated element cost or quantity.

For example, an original item: “Surface course: 10 mm, thin surface course system to Clause 000 35-mm thick passive stack ventilation (PSV) 65 in carriageway” with associated costs and quantities expressed in square meters, when reclassified into the standard WBS, requires costs and quantities to be expressed in cubic meters. The appropriate conversion factor, in this case, the thickness of the surface course (0.035 m [35 mm]) needs to be applied to the associated costs and quantities of the reclassified element. The example description, containing a number of numerical references, illustrates that selecting the appropriate conversion factor is non-trivial if the process is to be automated.

This study sought to develop an automated process to address these issues: The suggested approach first tokenizes the description at a word level and implements some data pre-processing.

Second, it uses one-hoot encoding to search relevant keywords that allow the algorithm to understand in which category each item should be located.

Finally, it provides predictions automatically to enhance and facilitate the accountant surveyor’s work automatically classifying three attributes: identifying the WBS code, identifying which items are in need of conversion and inferring the rate based on the text description, and identifying does items classified as “extra over” works.

3.1 | The input dataset

The input dataset was composed of 59,140 road infrastructure work items from submitted BOQs and a schedule of rates in the form of a structured spreadsheet. Each item was made up of 10 attributes as detailed in Table 1.

The example in Table 2 will give a flavor of the dataset.

A considerable number of the 59,140 descriptions were exactly repeated thus the number of unique items in the dataset was 17,452.

A total of 26 different units of measure were included, including length (in meters), areas (in square meters), time (in weeks), and weight in (tons), among others.

A total of 214 different WBS codes were present, irregularly distributed across categories.

TABLE 1 Attributes included in the dataset.

Attribute	Type	Description
Original data as submitted by the supply chain:		
project identifier number (PIN)	String	Project identifier
Description	String	Element description (a free text entry by the project team (cost consultant//surveyors etc.)
Unit of measure	Class	The unit of measurement used for the cost item (e.g., length (m), area (m^2), volume (m^3), and so forth
Series allocation	Class	Mid-level MMHW category (broad category)
Manually reclassified data:		
Work breakdown structure (WBS) code	Class	Classification code using the organization’s standard WBS
Subcategory name	String	Standard description of work item associated with the WBS code
Subcategory unit	Class	Unit of measure associated with the WBS code above. (May differ from the original unit of measure.)
Extra over	Boolean	Indicates whether the item is “extra over” or is a main cost item
Conversion	Double	Identifies a conversion factor, if any, to be applied to costs and quantities associated with the item resulting from a difference between the units of measurement used for the original data and those used for the reclassified WBS category

Abbreviation: MMHW, Method of Measurement for Highway Works.

Data on “extra over” items was similarly imbalanced with only 3264 (6%) out of a total of 59,140 unique rows classified as such.

Finally, 15,408 rows (26%) had been identified as requiring the application of a conversion factor.

3.2 | Implementation of the proposed automated approach

The suggested approach can be summarized into five steps that happen on a sequential basis:

3.2.1 | Step 1: Pre-processing

Description: This process automatically detects and fixes common errors such as typos or data quality issues, including:

TABLE 2 Sample values of the first five rows in the input dataset.

Unique ref	PIN	Item ref	Description	Unit	Series alloc.	WBS code	Subcat name	Subcat unit	Extra over	Conv.
6031916	603191	1500_07_030	Loop detection installation three lane…	No	1500	1500080	Loop detector install	No	False	1
6031918	603191	1200_03_190A	Minimum Visit charge roadmaking…	Sum	1200-300	1200-300060	Minimum lining visit	No	False	1
6031918	603191	1200_03_810	Minimum Lining Visit charge…	No	1200-300	1200-300060	Minimum lining visit	No	False	1
6031918	603191	1200_03_810	Minimum Lining Visit charge…	No	1200-300	1200-300060	Minimum lining visit	No	False	1
6031918	603191	1200_04_010	Bi-directional non-depressive…	No	1200-400	1200-400010	Road stud	No	False	1

Unit of measure: The original data included several typographical mistakes, presumably due to manual entry. Singular and plural units were also often used interchangeably, for example, “meter” and “meters.” There were also discrepancies in the use of capitalization, spelling, and inconsistent abbreviations such as “sq m” or sqm in place of “m^2.”

Description: The use of capital letters, special characters, and stop words was inconsistent. The stop words are a group of words composed of conjunctions, prepositions, articles, and adverbs that are generally filtered out before the application of NLP.

Challenges: The main complication found was coping with data variability; since the BOQ document has been manually written by different persons, they are prone to use different abbreviations or synonyms that identify the same concepts. To cope with this limitation, a thorough text analysis has been carried out to identify a list of the most common ambiguities.

Output: As a main result, the text descriptions are being pre-processed, and the descriptions are being tokenized at a word level to analyze the text at a word level.

3.2.2 | Step 2: Searching for keywords

Description: During this step, the algorithm takes an array of already tokenized words and implements one-hoot encoding to search for two groups of keywords: The first group contains 39 keywords with relevant information for identifying “extra over” cases and when an item needs conversion inside each description.

The second group is composed of 74 keywords that later will help the algorithm to identify the different WBS categories to which the item belongs.

The list of keywords that the algorithm searches for identifying extra over and conversion rates is as follows: “adjustment,” “material,” “install,” “maintain,” “excavation,” “supply,” “disconnect,” “testing,” “plan,” “establish,” “mill,” “exceed,” “additional,” “lieu,” “sign,” “traffic,” “carriageway,” “hard shoulder,” “management,” “removal,” “safety,” “extra,” “over,” “existing,” “additional,” “work,” “connect,” “night,” “dispose,” “thermoplastic,” “diagram,” “closure,” “millimetres,” “metre,” “sqm,” “litre,” “hour,” “PSV,” “km.”

Alternatively, the list keywords for classifying the WBS categories is: “bridge,” “expansion,” “joint,” “drain,” “service,” “surveys,” “protection,” “connection,” “surface,” “headwall,” “frame,” “drainage,” “reinstall,” “guttering,” “material,” “excavation,” “disposal,” “acceptable,” “unacceptable,” “topsoil,” “cable,” “communication,” “commissioning,” “equipment,” “works,” “ecologist,” “light,” “binder,” “tack coat,” “asphalt,” “reinstatement,” “pavement,” “over banding,” “planning,” “adjustment,” “cat man,” “piling,” “wall,” “cathodic protection,” “bollard,” “traffic sign,” “temporary road,” “minimum lining visit charge,” “road stud,” “information boards,” “provision,” “facilities,” “contractor,” “temporary,” “demolition,” “structures,” “item,” “removal,” “barrier,” “formwork,” “repair,” “testing,” “steel,” “establishment,” “plant,” “gate,” “maintenance,” “emergency,” “bifurcation,” “terminal,” “crash,” “transition,” “system,” “foundation,” “post,” “pedestrian,” “vehicle,” “other,” “concrete.”

Feature importance provides an indication of how important the occurrence of a given keyword is in making predictions of “extra over,” conversion or classifying the WBS categories. Although the feature importance can also be extracted indirectly in the SVM and SGD by comparing

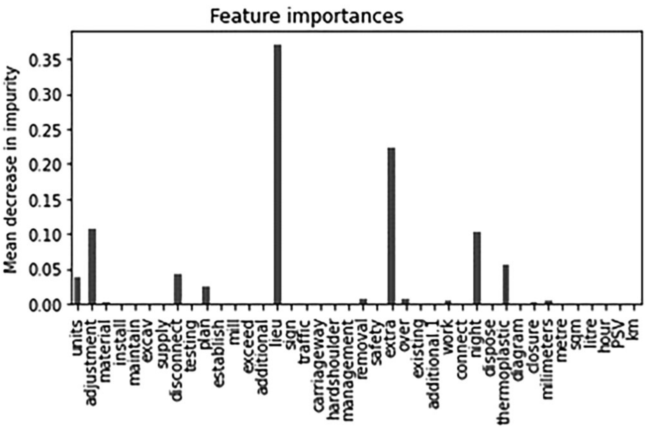

FIGURE 1 The feature importance of all the keywords found in the description using "extra over" as the class attribute.

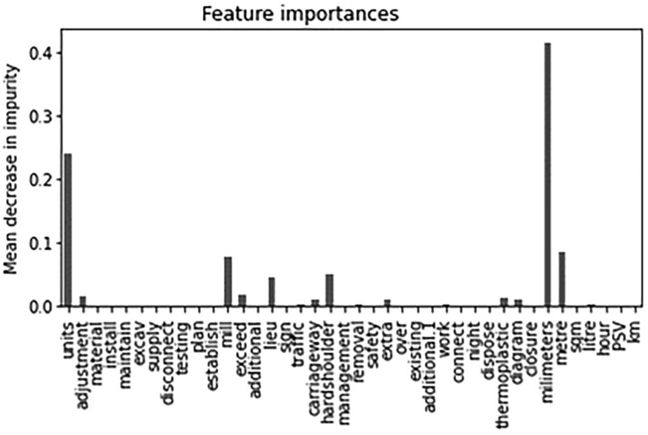

FIGURE 2 The feature importance of all the keywords found in the description using "conversion" as the class attribute.

the coefficients of their attributes, the feature importance of the random forest has been mainly considered due to its transparency and the better interpretability of its results.

Figure 1 shows the feature importance of keywords predicting when an item is an extra over. Second, Figure 2 shows the feature importance for binary classification predicting whether an item is in need of conversion. Finally, Figure 3 presents the feature important when predicting the WBS category codes using a different group of keywords.

Figure 1 highlights, unsurprisingly, that some keywords such as "extra" are very important. However, the algorithm also identified other, less obvious, relevant keywords, such as "disconnect," "plan," "establish," "night," and "thermoplastic."

To provide a practical example, the algorithm is able to infer that whenever the text description contains the word "night shift" or "night works" that implies that it is a work

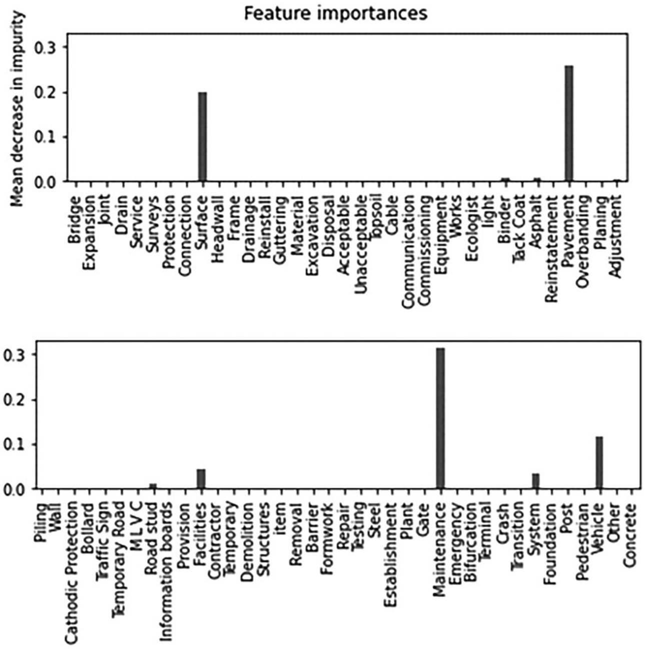

FIGURE 3 The feature importance of all the keywords found in the description when predicting the work breakdown structure (WBS) categories.

that was not originally planned and hence that should be classified as extra over.

Figure 2 highlights the importance of units of measure such as "millimeters" or "meters." In other words, it could be said that whenever the description specifies a unit of measure, that makes the algorithm infer that it is in need of conversion. The feature importance also highlighted some less obvious keywords such as "mill," "exceed," "hardshoulder," and "removal," The figures also demonstrate the difference in the importance of keywords for predicting one or another of the categories.

The results shown in Figure 3 should be taken with some skepticism, as it is possible to appreciate that the keywords targeting the most common WBS categories appear as more important.

Additionally, the keywords that appear more often in the text also increase their feature importance due to their variability. However, it can be inferred that the keywords that appear as less important are also helpful to classify the WBS categories that occur with less likelihood.

Challenges: The functionality of this step in itself is a challenge, as extracting relevant information out of text descriptions for classifying three different parameters is a very specific and hence an innovative job.

The two lists of the selected keywords are the result after carrying out a process of testing and refinement.

This process of refinement has been based on the feature importance of every single keyword when making the corresponding predictions and based on the differences in accuracy from the binary classifiers including and excluding the selected words. Additionally, the accountant surveyor's knowledge has been implemented to be taken into account in the choice of keywords.

Output: As a main result, two new datasets containing the following attributes are being generated, one of them will be used for "extra over" and conversion classification, whereas the second one will be used to classify the WBS categories:

The dataset for "extra over" and conversion classification contains the following attributes:

- A description identifier is used to link it with the input dataset.
- Thirty-nine Boolean attributes, one for each keyword indicating whether the description contains that keyword or not.
- A cleansed version of the original unit of measure from the input dataset.
- A Boolean attribute that indicates whether an item is "extra over" or not.
- An integer number indicates the conversion rate for that item. If the integer number is different than 1, it indicates that some conversion is needed for this item.

Alternatively, for the WBS code classification the generated dataset is as follows:

- A description identifier is used to link it with the input dataset.
- Seventy-four Boolean attributes, one for each keyword indicating whether the description contains that keyword or not.
- A cleansed version of the original unit of measure from the input dataset.
- A classification attribute indicating the WBS category that the item belongs to.

For clarification, Table 3 shows the values from the first five rows of the newly generated dataset to give a flavor of this output.

3.2.3 | Step 3: Making predictions

Description: This step makes use of three different ML classifiers that when applied to two different datasets provide predictions for three class attributes:

"Extra over": A Boolean attribute indicating whether an item is "extra over" or not.

"Conversion needed": Another Boolean attribute indicating whether the item needs conversion or not.

"WBS code": A classification attribute identifying the WBS category that each particular asset belongs to.

The three ML algorithms have been selected to carry out the classification task. The selection of the classifiers relies on their different way of working to establish a comparison, their transparency when generating the results, and their robustness:

1. *SVMs*—a supervised learning model originally developed for binary classification but also suitable for multiple classification and regression. The underpinnings of the algorithm rely on the maximal margin classifier, which is based on the hyperplane (Mayhua-López et al., 2015).

Despite the fact of being a classical algorithm, SVM was chosen as a classifier due to their proven accuracy when implemented in other construction sector projects, their memory efficiency, and last but not least their efficacy in handling scenarios in a high dimensional space as we have in the presented approach.

Some hyper-parameter Turing process has been carried out to maximize the accuracy of the model without overfitting it. As a main result of this process, the SVMs have been implemented with the following configuration:

- The kernel type used for the algorithm is the radial basis function.
- The kernel coefficient chosen is scale.
- The regularization parameter is 0.5.

2. *Random Forest*—It is an ensemble method for performing supervised classification or regression whose inner way of working consists of constructing a large number of decision trees when building the model (de Ville, 2013).

Random forest was selected for classification for two main reasons: First, due to their robustness, which ensures that their results will not vary meaningfully when implementing them in a different scenario. Second, their transparency allows for a better understanding of the extracted features and to infer why the algorithms are providing these predictions.

Likewise, the hyper-parameter Turing process for this algorithm indicated that a lower number of more complex

TABLE 3 Sample values of the dataset that were generated because of the second step.

Id	Units	Adjust.	Material	Install	Maint.	Excav.
1	False	False	False	False	False	False
2	False	False	False	False	False	False
3	False	False	False	False	False	False
4	False	False	False	False	False	False
5	False	True	False	False	False	False
	Litre	**Hour**	**PSV**	**Km**	**Extra over**	**Conv.**
…	False	False	False	False	False	True
…	False	False	False	False	False	True
…	False	False	False	False	False	True
…	False	False	False	False	False	True
…	False	False	False	False	False	True

trees outperforms other configurations. More specifically, the chosen configuration was as follows:

- The number of trees included in the forest is 600 to find a balance between efficiency and accuracy.
- The maximum number of features used for each tree would be 27.
- The maximum depth of each tree would be four.

3. *SGD*—an optimization algorithm usually combined with ML to find the best fit between predicted and current outputs. It consists of a simple and very efficient approach to fit linear classifiers and regressions under convex loss functions (Minoura et al., 2013).

The reason for picking the SGD algorithm relies on its inner way of working, which is completely different from the other two chosen classifiers and allows a more powerful better assessment when comparing the results of the three algorithms.

Despite the great accuracy achieved by this algorithm, the hyper-parameter Turing process showed in many cases the default values were the best configuration when applied to this scenario. As a result, the algorithm has been configured as follows:

- The loss function used would be a hinge.
- The penalty chosen would be "l2" left by default.

Challenges: The main issue found when predicting the "extra over" attribute was the fact that the algorithm was easily biased. The 95% of the instances presented a no extra over the asset that makes the models become easily biased and found difficulties predicting the cases where an "extra over" work occurs.

Alternatively, the WBS classification is not an easy task since the instances are irregularly distributed among a wide range of categories.

The way to tackle these issues consists of three main tasks: to choose the most suitable ML algorithms applied to this case, to implement feature engineering and extract useful information for classification, and to apply hyper-parameter Turing.

Selecting the correct ML algorithms is critical to achieve high accuracy and robust results. Due to the nature of the problem, only supervised learning algorithms were considered. Within this subset, the most simplistic algorithms such as k-nearest neighbours (KNN) and linear regression have been discarded due to their inner way of working.

As a last step, the algorithms were chosen from a subset after a process of testing and refinement depending mainly on their achieved accuracy, their efficiency, and their differences when calculating the results to make a richer comparison.

Second, a feature engineering process has been carried out as explained in detail in the previous step.

Finally, a process of feature engineering and hyper-parameter Turing has been performed assessing not only their impact on accuracy but also analyzing the instances where the chosen algorithm was wrong to ensure that it has not been overfitted.

Output: The main role of this step consists of automatizing the classification tasks of three attributes that were manually categorized by accountant surveyors. As a main result, the predictions of three different attributes are being provided: "Extra over," conversion, and the WBS category. These predictions will be used later for assessing the performance of the algorithms, for performing an automatic data classification, and for inferring misclassifications in the historically registered data.

3.2.4 | Step 4: Inferring the conversion factor

Description: In this step, the assets that need conversion have been taken as input, the main functionality of this step consists of inferring the conversion factor based on their text descriptions, emulating, and automatizing the work of the accountant surveyors.

To achieve this goal, a combination of some exploratory data analysis (EDA) to have a deeper knowledge of the input data and a set of meetings to fully understand the work of the accountant surveyors have been carried out.

Based on the findings of the EDA, the assets in need of conversion are being classified into six different types, and their conversion factor is being calculated by applying a set of rules:

CASE 1: When the "extra over" attribute is positive, the conversion factor would always take a special value to identify those cases and hence no further calculations are needed.

CASE 2: When the unit of measure for an asset is "week," the suggested conversion factor is 7, based on a simple analysis of the dataset (98% of instances had been manually assigned this conversion factor).

CASE 3: When the attribute unit of measure is "item," "meter," "square meter," "cubic meter," or "number." For these assets, the text descriptions are being tokenized at a word level seeking words completely composed of numbers.

In the case where more than one number is being identified, the algorithm extracts the quantities and picks the most appropriate number depending on the unit of measure specified in the description and establishes priority rules.

The text analysis of the algorithm is based on the search for units of measures or specific keywords appearing in the words immediately after the extracted numbers.

For clarification, a practical example is described. In a fictitious description, "the asset is 100 meters thick, 20 meters high, and 30 tons of other material." The model would firstly extract the words 100, 20 and 30 as three possible conversion factors.

Second, it would search for units of measure in the words immediately before and after. In this case, it will infer that there are 100 m and 20 m, and it will discard 30 tons presuming that the unit of measure "tons" does not match with the corresponding measurement for this asset.

Finally, the approach establishes different priorities and based on the EDA that has been previously carried out, the algorithm identifies that the thickness is the metric that determines the conversion factor and not the height since that is the general behavior for the subset of the historically classified assets.

CASE 4: The unit of measure is "liter." In this case, a similar process to the Case 3 is carried out. In this case, the priority would be the numbers that include the unit of measure "liter" and its abbreviations "Ltr" and "l."

CASE 5: The unit of measure is "ton." Likewise, a similar process is implemented to identify the word "ton" and its abbreviations: "tn" and "t."

CASE 6: If the unit of measure is: "day," "shift," "night," or "sum." The algorithm would infer that no conversion is needed, and hence the asset has been wrongly classified in the previous step. The EDA that has been previously carried out indicated that over 90% of the assets containing these units of measure do not need conversion.

Challenges: The role of this step is a very unique task aiming to transform the accountant surveyor's knowledge into a set of simplified rules that produce an equivalent output. The novelty of the task and the main aim of it presented difficulties when designing the conversion rules.

Additionally, the descriptions were manually included by different workers, which implies a great variability in them. As a result of that, the algorithm identifies that "meter," "meters," "m," or "ms" should be considered the same words. When establishing the priorities in some cases the algorithm is also able to identify specific synonyms.

Output: As a main result, the numeric conversion factor for all those cases in need of conversion is being extracted enhancing the role carried out by the machine classifiers in the previous step.

3.2.5 | Step 5: Detecting inconsistencies in the historically registered assets

Thus far, a fully automatized method has been presented to perform the tasks of data wrangling and a reliable classification. However, the method has also been executed and its results have been contrasted with accountant surveyors to detect inconsistencies in the historically registered data. Obtaining a new functionality as a low-hanging fruit.

For clarification, two detected inconsistencies in the dataset are provided:

In the first example, the following test data were submitted to the classification system:

- Description: "Extruded Thermoplastic Lining (Lining Material Only)"
- Unit of measure: tone
- Series allocation: 1200–400

The system provided the following classification:

- Classified as no "extra over."
- Conversion needed with a conversion rate of 166.67.

An inconsistency has been detected: The description is not specific enough. The algorithm or any human manually carrying out the task cannot infer if the item needs conversion with that specific rate with the given description.

In the second example, the following test data were submitted:

- Description: "Servicing of all mobile welfare facilities for Task Order value not exceeding £XXX."
- Unit of measure: week.
- Series allocation: 100
- Classified as no "extra over."
- It needs conversion with a conversion rate of 5.
- The original unit of measure of the item is week.
- The unit of measure of the category for the item is a day.

In the second case, some incoherencies have been detected. The description is not specific enough to allow classification for the system or for any human manually performing the classification task. Additionally, it could be inferred that a typo has occurred classifying the week as 5 days instead of 7.

3.3 | Tools and materials

The presented solution has been programmed using Anaconda Navigator, Jupyter Notebook version 6.5.2 and Python programming language version 1.8. Implementing the corresponding libraries. From the libraries used it is important to highlight "nltk" implementing NLP functionalities and "scikit learn" to implement some ML capabilities.

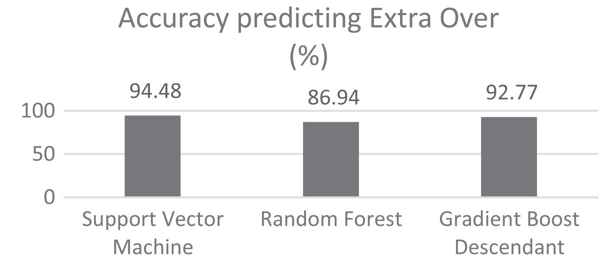

FIGURE 4 Percentage of instances classifying correctly "extra over" works in all algorithms.

4 | RESULTS

The assessment of the results provided by the suggested methods is composed of three main parts:

First, the results provided in step three generated by the ML algorithms when predicting the three class attributes will be broken down and reviewed.

Second, the accuracy of the algorithm inferring the conversion factor based on the text descriptions will be commented out.

Finally, the benefits of the method in terms of efficiency will be calculated. For that, a comparison of the same scenario with and without the implementation of the suggested method will be established.

4.1 | The results of the classifiers

To validate the accuracy of the ML models, the cross-validation method (Porta, 2014) with 10 folds was implemented separately for each of the three class attributes:

When predicting the "extra over" items (see Figure 4), the SVM algorithm managed to classify correctly 94% of the instances, whereas gradient boost descendent correctly classified 92% of them in contrast with random forest, which only obtained 87% accuracy.

Breaking down the results classifying the "extra over" attribute, it is important to take into account that 94.48% of the instances consist of items classified as no "extra over," whereas the remaining 5.52% are classified as extra over, despite the fact that the configurations of the algorithms were adjusted to avoid this issue. Unavoidably, the algorithms are more prone to correctly predict the no "extra over" works.

More accurately, the confusion matrix indicates that the SVM was able to predict correctly 26 instances classified as "extra over," out of 3264. Alternatively, the random forest correctly predicted "extra over" works 223 times, whereas the gradient boost descendant (GBD) was right 35 times.

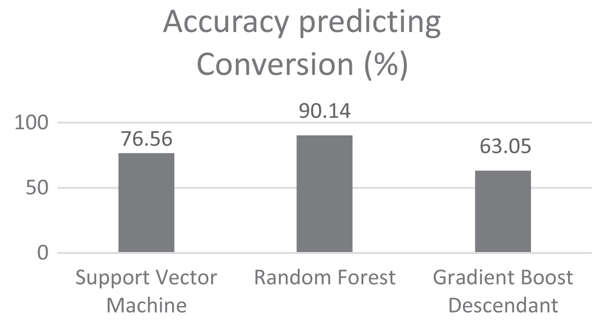

FIGURE 5 Percentage of instances correctly classified predicting when an item needs conversion in the three algorithms.

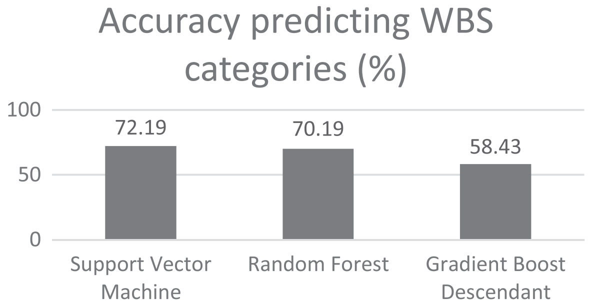

FIGURE 6 Percentage of instances classified correctly when predicting WBS categories.

Hence, even though the random forest presented the lowest accuracy, it is also the most unbiased algorithm. That is due to its inner way of working. The algorithm is based on classifying instances and establishing internal rules based on the data patterns and that feature makes it more robust in scenarios where the algorithms are prone to be overfitted.

Regarding the predictions detecting whether an item is in need of conversion. An accuracy summary has been provided in Figure 5, indicating the percentage of instances classified correctly for this class attribute. In this case, random forest was the most accurate with over 89% of the instances classified correctly, followed by SVM and GBD, which only achieved 61% accuracy.

Taking a closer look at the dataset, it is possible to infer that this class attribute presents a less uneven distribution with 76.57% of the occurrences classified as no conversion needed, whereas the remaining 23.43% needed conversion.

A total of 13,856 instances were in need of conversion, and analyzing only this subset, SVM classified correctly 309, the random forest was right 1507 times, whereas SGD correctly detected the conversion 153 times.

The slightly better accuracy of the random forest algorithm in a less unbiased scenario reassures the idea that it is a more robust algorithm, and hence that would be the recommended choice for the system.

Finally, Figure 6 illustrates the percentage of instances classified correctly by predicting all the category codes using a different group of keywords and their allocation series.

The SVM became the most accurate algorithm with 72% accuracy and the random forest with 70% accuracy, being meaningfully higher than the gradient boost descendant, which only classified correctly the 58% of the instances.

Taking a closer look at the results, it is important to take into consideration that the algorithms used a dataset created by different attributes. Additionally, the problem to be faced is different now since the algorithms do not aim to perform a binary classification but multiple classifications identifying 214 different WBS codes irregularly distributed on the dataset.

4.2 | The results inferring the conversion rates

For the validation of the results when predicting the conversion rates, the algorithm takes as inputs all the items in need of conversion, which are a total of 15,408 as identified in the original dataset. To assess the accuracy of results, the algorithm runs iteratively through each one of the instances, and its results are compared with the historical values, stored in the dataset.

The results show that the algorithm was able to predict the conversion rate correctly by analyzing the text in 13,449 (87.28%) of cases. However, this value should only be taken as guidance. The fact that there is a discrepancy between the predictions and the real historical values does not necessarily imply that the algorithm was not robust enough since in some cases, some inconsistencies in the original manual classification of the data were found by inspection.

4.3 | Measuring the benefits of the method

The main aim of the method consists of providing an automatic classification of three different attributes that were manually classified before. To demonstrate the benefits of the suggested solution in terms of efficiency, an estimation will be made calculating the amount of time that would take for an accountant surveyor to classify and register the information for each asset in the same scenario:

Without using the method:

The accountant surveyor would need to register the following information:

- The description of the asset: (2 min)

- To include the reference, the scheme identifier, and the unit of measure (1 min)
- To calculate the WBS category that it belongs to (1 min)
- To binary classify the conversion including, if necessary, the conversion rate, and identify the “extra over” works (1 min).

All in all, it would take the accountant surveyor 5 min to correctly classify and register each asset, multiplied by the number of assets 59,140, which would make a total of 4928 h and 20 min.

Using the method:

The accountant surveyor would need to register the following information:

- The description of the asset: (2 min)
- To include the reference, the scheme identifier, and the unit of measure (1 min)
- To review the automatic classifications of the WBS codes, “extra over” works, and the conversion rates (30 s).

All in all, it would take the same person 3 min and 30 s, which means the time for registering the same asset has been reduced by 30%, multiplied by the total number of assets 59,140, and that would make a total of 3449 h and 50 min. In other words, we could say that the implementation of the method has saved the construction company 1478 h of work.

Additionally, the construction company obtains additional benefits. The automatic classifications make a more reliable system and are less prone to subjectivities and the generated data take less time to be usable due to its efficiency. In addition to that, the system can also be used to detect inconsistencies and misclassifications as demonstrated in the fifth step of the approach.

5 | DISCUSSION

This paper presented a full-fledged methodology for automating the identification of inconsistencies and classification of infrastructure cost data. This research paves the way for harnessing artificial intelligence (AI) to improve data mining in the construction and infrastructure sector, especially when the collected data are prone to inconsistencies. This study provides a step change in comparison to previous studies, where the application of AI has been constrained to cost estimation and prediction, with little focus on automating the mining and pre-processing phases (Elmousalami, 2020; Pham et al., 2021; Tayefeh Hashemi et al., 2020). So far, little work has addressed the challenges of improving data quality, yet it was stressed that data quality and usability are key inhibitors of AI data analytics and cost intelligence in the construction sector (Thomas & Bowman, 2021). Furthermore, the proposed novel approach for automated classification of infrastructure cost data into a consistent format outperformed previous attempts, such as Chen et al. (2019) and Martínez-Rojas (2015).

In addition, the research provided a comprehensive approach through a combination of ML and expert knowledge representation and applying data science, well beyond previous work, such as Soibelman et al. (2008) and Matthews et al. (2022), which relied heavily on text analytics.

The results show a high level of accuracy (i.e., 94%) when predicting inconsistencies related to “extra over” works, in contrast to previous studies, which also used SVM for performing classification between different model disputes resolutions (i.e., mediation, arbitration, litigation, negotiation, and administrative appeals), obtaining only 77.04% accuracy (Chou et al., 2013).

In other fields of application, other studies have also implemented SVM for binary classification obtaining 91.51% accuracy in the case of the proximal SVM, 91.73% when using ε-proximal SVM, and 91.93% when using the cognitive SVM (Zhu et al., 2015). In comparison to mining infrastructure and construction data, the data in these domains of application tend to be better structured and thus pose fewer challenges, and yet it has achieved a lower level of accuracy.

However, it is also important to acknowledge some limitations of the proposed approach. For example, the random forest algorithm obtained 90% accuracy when an item needs conversion. Other comparable studies, using the same algorithm applied to construction projects obtained an R-square ranging from 0.691 to 0.871, which implies that the algorithm was able to explain from 69% to 87% of the variation that occurred in the predicted feature (Cha et al., 2020).

Random forest (RF) was used in other studies, outside infrastructure, and construction, to perform a binary classification, obtaining, for example, 87.5% accuracy for the prediction of the gender of salmons (Guragain et al., 2022), 78% accuracy to foresee whether the water contains nitrate, and 74% for detecting iron and arsenic in the water (Tesoriero et al., 2017).

Finally, for inferring the conversion rate, the solution obtained 87% accuracy when predicting the conversion rate and 72% accuracy when classifying between the 214 WBS codes.

It is important to stress that the predictions were generated by harnessing data science when predicting the conversion rates and by implementing SVM for classifying the WBS codes.

Various approaches in construction management applied regression analysis. For example, 75% accuracy was obtained when applying logistic regression for predicting a contractor´s performance (Wong, 2004). However, when predicting construction costs using multiple regression techniques, a mean absolute percentage error of 19.3% was obtained, implying an accuracy of 80.7% (Lowe et al., 2006).

The results of our proposed approach demonstrate that it is possible to build a fully automated methodology to harness advances in text analysis and ML to emulate an expert's knowledge, when classifying items as a part of the bill of quantities document.

The used dataset contained 59,140 historically registered items, which implies that it is a dataset with a reasonably large and reliable quantity and contained results that support the robustness of the proposed approach.

6 | CONCLUSION

The research reported in this paper demonstrated the potential of robust ML-based approaches in infrastructure cost data mining, more specifically for facilitating automated data pre-processing and classification to consistent data structures. The work has shown promising results in the development of an integrated approach using three different classifiers, word tokenization and one-hoot encoding, and text analytics to infer the correct conversion rates.

The presented approach shows a validated method to support automation and detection of inconsistencies in data for infrastructure cost data analysis and benchmarking.

Based on the findings of SVM, it is possible to infer that the algorithm was able to achieve a greater accuracy although the random forest algorithm would be the recommended choice. By analyzing the feature importance, extracted from the random forest algorithm, it was possible to determine their accuracy when predicting whether an item is in need of conversion. This depended on the occurrence of keywords related to the unit of measure, such as meter or tonne.

Following cross-validation, the ML algorithms demonstrated that the solution was robust and contained a high level of accuracy, when applied to a real scenario of infrastructure projects. The solution was successful at detecting inconsistencies based on the unit of measure and correcting identified inconsistencies. The preferred algorithm correctly classified 87% of the cases for “extra over” and detected 90% of the cases for whether an item needed conversion. The approach was also able to correctly predict the conversion factor with 87% accuracy and could distinguish between the 214 different WBS categories with 72% accuracy.

The proposed AI-based approach has the potential to pave the way for the development of robust automated cost data classification systems in the infrastructure sector.

ACKNOWLEDGMENTS

This research was supported by the UK Department for Transport and received funding under the Innovate UK project “Transport Infrastructure Efficiency Strategy Living Lab (45382).”

REFERENCES

Adeli, H., & Karim, A. (1997). Scheduling/cost optimization and neural dynamics model for construction. *Journal of Construction Management and Engineering*, *123*(4), 450–458.

Adeli, H., & Kim, H. (2001). Cost optimization of composite floors using the neural dynamics model. *Communications in Numerical Methods in Engineering*, *17*, 771–787.

Ahiaga-Dagbui, D. D., & Smith, S. D. (2012). Neural networks for modelling the final target cost of water projects. In S. D. Smith (Ed.), *Proceedings of the 28th annual ARCOM conference* (Vol. 2, no 1, pp. 307–316). Association of Researchers in Construction Management.

Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings*, *147*, 77–89. https://doi.org/10.1016/J.ENBUILD.2017.04.038

Akinosho, T. D., Oyedele, L. O., Bilal, M., Ajayi, A. O., Delgado, M. D., Akinade, O. O., & Ahmed, A. A. (2020). Deep learning in the construction industry: A review of present status and future innovations. *Journal of Building Engineering*, *32*, 101827.

Bodendorf, F., Merkl, P., & Franke, J. (2021). Artificial neural networks for intelligent cost estimation—A contribution to strategic cost management in the manufacturing supply chain. *International Journal of Production Research*, *60*(21), 6637–6658.

Bottou, L. (2014). From machine learning to machine reasoning: An essay. *Machine learning*, *94*(2), 133–149. https://doi.org/10.1007/S10994-013-5335-X/FIGURES/13

Cha, G.-W., Moon, H. J., Kim, Y.-M., Hong, W.-H., Hwang, J.-H., Park, W.-J., & Kim, Y.-C. (2020). Development of a prediction model for demolition waste generation using a random forest algorithm based on small datasets. *International Journal of Environmental Research and Public Health*, *17*(19), 6997. https://doi.org/10.3390/IJERPH17196997

Chen, D., Hajderanj, L., & Fiske, J. (2019). Towards automated cost analysis, benchmarking and estimating in construction: A machine learning approach. *Multi Conference on Computer Science and Information Systems, MCCSIS 2019—Proceedings of the International Conferences on Big Data Analytics, Data Mining and Computational Intelligence 2019 and Theory and Practice in Modern Computing 2019*, Porto, Portugal (pp. 85–91). https://doi.org/10.33965/bigdaci2019_201907l011

Cheng, M.-Y., & Roy, A. F. V. (2010). Evolutionary fuzzy decision model for construction management using support vector machine. *Expert Systems with Applications*, *37*(8), 6061–6069. https://doi.org/10.1016/J.ESWA.2010.02.120

Cheng, M.-Y., & Wu, Y.-W. (2009). Evolutionary support vector machine inference system for construction management. *Automation in Construction*, *18*(5), 597–604. https://doi.org/10.1016/J.AUTCON.2008.12.002

Chou, J.-S., Cheng, M.-Y., & Wu, Y.-W. (2013). Improving classification accuracy of project dispute resolution using hybrid artificial intelligence and support vector machine models. *Expert Systems with Applications*, *40*(6), 2263–2274. https://doi.org/10.1016/J.ESWA.2012.10.036

de Ville, B. (2013). Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*(6), 448–455. https://doi.org/10.1002/WICS.1278

Elmousalami, H. H. (2020). Artificial intelligence and parametric construction cost estimate modeling: State-of-the-art review. *Journal of Construction Engineering and Management*, *146*(1), 03119008.

Florez-Perez, L., Song, Z., & Cortissoz, J. C. (2022). Using machine learning to analyze and predict construction task productivity. *Computer-Aided Civil and Infrastructure Engineering*, *37*(12), 1602–1616.

GDQH. (2021). *Hidden costs of poor data quality*. Government Data Quality Hub. https://www.gov.uk/government/news/hidden-costs-of-poor-data-quality

Gorse, C., Johnston, D., & Pritchard, M. (2012). *Dictionary of construction, surveying and civil engineering* (1st ed.). Oxford University Press.

Guragain, P., Båtnes, A. S., Zobolas, J., Olsen, Y., Bones, A. M., & Winge, P. (2022). IIb-RAD-sequencing coupled with random forest classification indicates regional population structuring and sex-specific differentiation in salmon lice (*Lepeophtheirus salmonis*). *Ecology and Evolution*, *12*(4), e8809. https://doi.org/10.1002/ECE3.8809

ICMS. (2021). *The international cost management standard* (3rd ed.). ICMS Coalition. https://icmscblog.files.wordpress.com/2021/11/icms_3rd_edition_final.pdf

Ji, S.-H., Ahn, J., Lee, H.-S., & Han, H. (2019). Cost estimation model using modified parameters for construction projects. *Advances in Civil Engineering*, *2019*, 8290935.

Langroodi, A. K., Vahdatikhaki, F., & Doree, A. (2021). Activity recognition of construction equipment using fractional random forest. *Automation in Construction*, *122*, 103465. https://doi.org/10.1016/J.AUTCON.2020.103465

Lin, J.-R., Hu, Z.-Z., Zhang, J.-P., & Yu, F.-Q. (2016). A natural-language-based approach to intelligent data retrieval and representation for cloud BIM. *Computer-Aided Civil and Infrastructure Engineering*, *31*(1), 18–33.

Liu, P., Xie, M., Bian, J., Li, H., & Song, L. (2020). A hybrid PSO–SVM model based on safety risk prediction for the design process in metro station construction. *International Journal of Environmental Research and Public Health 2020*, *17*(5), 1714. https://doi.org/10.3390/IJERPH17051714

Lowe, D. J., Emsley, M. W., & Harding, A. (2006). Predicting construction cost using multiple regression techniques. *Journal of Construction Engineering and Management*, *132*(7), 750–758. https://doi.org/10.1061/(ASCE)0733-9364(2006)132:7(750)

Martinez, E., Marcos, A., Al-Kassir, A., Jaramillo, M. A., & Mohamad, A. A. (2012). Mathematical model of a laboratory-scale plant for slaughterhouse effluents biodigestion for biogas production. *Applied Energy*, *95*, 210–219. https://doi.org/10.1016/J.APENERGY.2012.02.028

Martínez-Rojas, M., Marín, N., & Vila, M. A. (2015). An approach for the automatic classification of work descriptions in construction projects. *Computer-Aided Civil and Infrastructure Engineering*, *30*(12), 919–934.

Matthews, J., Love, P. E. D., Porter, S. R., & Fang, W. (2022). Smart data and business analytics: A theoretical framework for managing rework risks in mega-projects. *International Journal of Information Management*, *65*, 102495.

Mayhua-López, E., Gómez-Verdejo, V., & Figueiras-Vidal, A. R. (2015). A new boosting design of support vector machine classifiers. *Information Fusion*, *25*, 63–71. https://doi.org/10.1016/J.INFFUS.2014.10.005

Minoura, K., Tamura, S., & Hayamizu, S. (2013). Probabilistic expression of polynomial semantic indexing and its application for classification. *Pattern Recognition Letters*, *34*(13), 1485–1489. https://doi.org/10.1016/J.PATREC.2013.05.009

MMHW. (2009). *Method of measurement for highway works. manual of contract documents for highway works (section 1,2,3)*. www.standardsforhighways.co.uk/tses/attachments/1e56dda3-cc10-4588-90be-a1e9a8a7870a

Pham, T. Q. D., Le-Hong, T., & Tran, X. V. (2021). Efficient estimation and optimization of building costs using machine learning. *International Journal of Construction Management*, *23*(5), 909–921.

Porta, M. (2014). *Cross-validation* (6th ed.). Oxford University Press.

Ramos-Bernal, R. N., Vázquez-Jiménez, R., Cantú-Ramírez, C. A., Alarcón-Paredes, A., Alonso-Silverio, G. A., Bruzón, A. G., Arrogante-Funes, F., Martín-González, F., Novillo, C. J., & Arrogante-Funes, P. (2021). Evaluation of conditioning factors of slope instability and continuous change maps in the generation of landslide inventory maps using machine learning (ML) algorithms. *Remote Sensing 2021*, *13*(22), 4515. https://doi.org/10.3390/RS13224515

Rico-Juan, J. R., Gallego, A. J., & Calvo-Zaragoza, J. (2019). Automatic detection of inconsistencies between numerical scores and textual feedback in peer-assessment processes with machine learning. *Computers & Education*, *140*, 103609. https://doi.org/10.1016/J.COMPEDU.2019.103609

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, *34*(1), 1–47. https://doi.org/10.1145/505282.505283

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

Soibelman, L., Wu, J., Caldas, C., Brilakis, I., & Lin, K.-Y. (2008). Management and analysis of unstructured construction data types. *Advanced Engineering Informatics*, *22*(1), 15–27. https://doi.org/10.1016/j.aei.2007.08.011

Tayefeh Hashemi, S., Ebadati, O. M., & Kaur, H. (2020). Cost estimation and prediction in construction projects: A systematic review on machine learning techniques. *SN Applied Sciences*, *2*, 1703. https://doi.org/10.1007/s42452-020-03497-1

Tesoriero, A. J., Gronberg, J. A., Juckem, P. F., Miller, M. P., & Austin, B. P. (2017). Predicting redox-sensitive contaminant concentrations in groundwater using random forest classification. *Water Resources Research*, *53*(8), 7316–7331. https://doi.org/10.1002/2016WR020197

Thomas, E., & Bowman, J. (2021). *Construction data and analytics report: Harnessing the data advantage in construction*. Autodesk and FMi. https://construction.autodesk.com/resources/fmi-construction-data-report/?utm_medium=press-release&utm_source=blog&utm_campaign=fmi2021&utm_region=global

Wang, Z., Zhang, Y., Mosalam, K. M., Gao, Y., & Huang, S.-L. (2022). Deep semantic segmentation for visual understanding on construction sites. *Computer-Aided Civil and Infrastructure Engineering*, *37*(2), 145–162.

Wong, C. H. (2004). Contractor performance prediction model for the United Kingdom construction contractor: Study of logistic regression approach. *Journal of Construction Engineering and Management*, *130*(5), 691–698. https://doi.org/10.1061/(ASCE)0733-9364(2004)130:5(691)

Wu, C., Li, X., Guo, Y., Wang, J., Ren, Z., Wang, M., & Yang, Z. (2022). Natural language processing for smart construction: Current status and future directions. *Automation in Construction*, *134*, 104059. https://doi.org/10.1016/J.AUTCON.2021.104059

Zhao, X., Yeoh, K. W., & Chua, D. K. H. (2020). Extracting construction knowledge from project schedules using natural language processing. In K. Panuwatwanich & C. H. KO (Eds.), *The 10th international conference on engineering, project, and production management*. Lecture notes in mechanical engineering (pp. 197–211). https://doi.org/10.1007/978-981-15-1910-9_17

Zhu, G., Huang, D., Zhang, P., & Ban, W. (2015). ε-Proximal support vector machine for binary classification and its application in vehicle recognition. *Neurocomputing*, *161*, 260–266. https://doi.org/10.1016/J.NEUCOM.2015.02.035

Zou, Y., Kiviniemi, A., & Jones, S. W. (2017). Retrieving similar cases for construction project risk management using Natural Language Processing techniques. *Automation in Construction*, *80*, 66–76. https://doi.org/10.1016/J.AUTCON.2017.04.003

How to cite this article: Dopazo, D. A., Mahdjoubi, L., Gething, B., & Mahamadu, A.-M. (2024). An automated machine learning approach for classifying infrastructure cost data. *Computer-Aided Civil and Infrastructure Engineering*, *39,* 1061–1076. https://doi.org/10.1111/mice.13114