Journal of Infrastructure, Policy and Development 2024, 8(5), 2801.
https://doi.org/10.24294/jipd.v8i5.2801

Article

Comparing data mining methods for predicting cost construction projects: A case study of cost management datasets from Thailand

Tanayut Chaitongrat[1], Kridtsada Janthachai[1], Wuttipong Kusonkhum[2], Paranee Boonchai[3,*], M. Faisi Ikhwali[4], Mathinee Khotdee[1]

[1] Construction and Project Management Center, Faculty of Architecture Urban Design and Creative Arts, Mahasarakham University, Mahasarakham 44150, Thailand
[2] Department of Civil Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen 40002, Thailand
[3] Faculty of Tourism and Hotel Management, Mahasarakham University, Mahasarakham 44150, Thailand
[4] Department of Environmental Engineering, Faculty of Science and Technology, Universitas Islam Negeri Ar-Raniry Banda Aceh, Banda Aceh 23111, Indonesia
*** Corresponding author:** Paranee Boonchai, Paranee.b@msu.ac.th

CITATION

Chaitongrat T, Janthachai K, Kusonkhum W, et al. (2024). Comparing data mining methods for predicting cost construction project: A case study of cost management datasets from Thailand. Journal of Infrastructure, Policy and Development. 8(5): 2801. https://doi.org/10.24294/jipd.v8i5.2801

ARTICLE INFO

Received: 5 September 2023
Accepted: 8 October 2023
Available online: 12 April 2024

Abstract: This research examines three data mining approaches employing cost management datasets from 391 Thai contractor companies to investigate the predictive modeling of construction project failure with nine parameters. Artificial neural networks, naive bayes, and decision trees with attribute selection are some of the algorithms that were explored. In comparison to artificial neural network's (91.33%) and naive bays' (70.01%) accuracy rates, the decision trees with attribute selection demonstrated greater classification efficiency, registering an accuracy of 98.14%. Finally, the nine parameters include: 1) planning according to the current situation; 2) the company's cost management strategy; 3) control and coordination from employees at different levels of the organization to survive on the basis of various uncertainties; 4) the importance of labor management factors; 5) the general status of the company, which has a significant effect on the project success; 6) the cost of procurement of the field office location; 7) the operational constraints and long-term safe work procedures; 8) the implementation of the construction system system piece by piece, using prefabricated parts; 9) dealing with the COVID-19 crisis, which is crucial for preventing project failure. The results show how advanced data mining approaches can improve cost estimation and prevent project failure, as well as how computational methods can enhance sustainability in the building industry. Although the results are encouraging, they also highlight issues including data asymmetry and the potential for overfitting in the decision tree model, necessitating careful consideration.

Keywords: contractor costs; data mining; neural networks; naïve bayes; decision tree

1. Introduction

The field of construction is very data-intensive and is rapidly expanding in terms of data development and collecting. The construction sector has transitioned into a digital era marked by an extraordinary growth in data volume in accordance with technological improvement (You et al., 2019). Data mining is a powerful tool to automatically uncover hidden knowledge from enormous and complicated data housed in databases, data warehouses, or other information repositories. It is a new and promising technology (Berry et al., 2004; Hui et al., 2000).

A significant volume of research on data mining applications in the construction sector has been produced in recent years. Several articles have examined data mining applications in the construction sector in this area. For instance, Bilal et al. (2016)

conducted a survey of the literature on the use of big data approaches in the construction sector. However, this study only analyzed a few works on data mining applications in the construction industry rather than providing a thorough introduction to the field. Data mining and its benefits for boosting performance in the Nigerian construction sector were explored by Aghimien et al. (2018). Although there is no global perspective, this study only focused on Nigerian data mining use in the construction industry.

In the realm of construction initiatives worldwide, with particular emphasis on Thailand, the recurring phenomena of cost escalation and project procrastination present intricate challenges. Such complexities compromise not only the financial health of construction entities but also pose substantial repercussions to the broader economic landscape, considering the influential nature of the construction sector (Dlamini et al., 2021). Traditional mechanisms for cost administration, while pivotal to project oversight, specifically within the domains of engineering and construction frequently exhibit inadequacies (Holm et al., 2021). Existing comprehension and methodologies for perpetuating cost regulation are marred by a lack of precision and habitual disregard for circumstantial and project-centric influences, thereby diluting the effectiveness of cost containment (Hansen et al., 2021; Gyadu-Asiedu et al., 2021).

Preceding scholarly endeavors have accentuated the correlation between effective cost administration and successful realization of construction ventures (Hoseini et al., 2020). Factors such as amendments in design schematics, lapses in site administration, deferred compensations by project proprietors, and capricious pricing of materials, wield a detrimental impact on cost efficacy (Faten Albtoush et al., 2020). Notwithstanding the indispensability of continual cost supervision for the smooth functioning of construction enterprises, contemporary stratagems frequently neglect the facet of resource streamlining for profit amplification (Borovskikh et al., 2021).

A comprehensive analysis of the scholarly landscape, including but not limited to the works of Chen et al. (2019) reveals a palpable deficiency in the exploration and scrutiny of the bearing of these cost administration variables on the unsuccessful execution of construction partnering ventures, especially within an economic context. This scholastic void gains pronounced urgency given the escalating global advocacy for sustainable practices and the profitable potential of environmentally conscious construction projects (Aisheh, 2021). In an endeavor to bridge these lacunae, the present research proposes a groundbreaking methodology employing data mining algorithms to prognosticate the collapse of construction projects, with particular attention to cost administration predicaments. The adopted data analytics tools, namely, Artificial Neural Network (ANN), Naïve Bayes (NB), and Decision Tree (DT), have been chosen based on their established track record in solving classification conundrums, specifically in the realm of predicting business failures. These are popularly tools and high accuracy with construction studies with artificial technologies (Kusonkhum et al., 2022; Sweis et al., 2008).

Finally, this study applies these techniques to primary data amassed from 391 contracting firms in Thailand, thereby yielding distinct, context-oriented findings. Through an examination of the relative performance of these data mining techniques in anticipating construction project failure, this study endeavors to formulate a predictive framework for industry failure in the construction realm, thereby fortifying

cost administration tactics. The fulfillment of this endeavor would serve not only to redress the scholarly deficiencies identified, but also to bestow invaluable, empirically backed insights upon construction firms, project overseers, and policy framers, thereby enabling them to make enlightened, data-driven decisions. This, in turn, will catalyze the incorporation of sustainable practices into the construction industry.

2. Project failure in construction management

It has been determined that the failure of construction projects is a highly frequent occurrence in Jordan, with theoretically lucrative plans typically turning out to be costly and loss-making ventures. This is disastrous for the owner as well as the contractor because future contracts may become less confident between them if the current project fails to live up to expectations. For the following reasons, identifying the root causes of project failure has proven to be a difficult challenge for academic scholars and professionals: First, there is a lack of consensus over the definition of project failure, leading to a lack of clarity in the project management (PM) literature. A second difficulty is caused by the likelihood that the causes of failure might vary in relation to the kind of project being studied (Sweis et al., 2008; Larson et al., 2013).

However, the difficulty of any project is in making it function and be successful within the Triple Constraint, which consists of the budget, time, and scope. A project's three components need to be in harmony with one another. Simultaneously, the other two parts will be impacted when one of them is restricted or stretched. The project manager must have a thorough understanding of each of the three components for there to be balance between them. This study acknowledges this controversy and defines project failure as going beyond budget, beyond schedule, and beyond scope in addition to not being able to achieve the functional requirements of the project as understood by stakeholders (Trost et al., 2003).

Numerous studies revealed that a variety of issues, including late decision-making, client-initiated modifications, and subpar site management and supervision, consistently contributed to construction project overbudgets (Trost et al., 2003). Many academics agree that a project has most likely failed when it takes longer than expected to complete, goes over budget, or produces results that don't meet predetermined performance metrics (Mohd et al., 2011). Moreover, according to some academics, these kinds of projects usually don't meet the needs of their consumers (Gündüz et al., 2013).

3. Data mining technique

A substantial portion of relevant scholarship has embarked on the journey of devising models purposed to foretell business failure probabilities, yet a minimal faction of these studies has endeavored into a comparative evaluation of the accuracies of these prognostic models. Our research intends to reconcile this gap by creating a predictive apparatus that utilizes data mining techniques such as Artificial Neural Network (ANN), Naïve Bayes (NB), and Decision Tree (DT) (Chamidah et al., 2020). The projected yields of this exploration are to render empirical substantiation pertaining to the consequences of variable input selection on predictive performance

over diverse periods, and to proffer instrumental inferences towards refining the precision of predicting the failure of construction businesses.

Multiple initiatives have been taken within the construction industry realm to devise model's adept at prefiguring business failures. With regards to predictive methodologies, a wide spectrum of strategies has been employed, spanning ANN, DT, Support Vector Machine (SVM), logistic regression (logit), and Multivariate Discriminant Analysis (MDA), among others. Albeit, a universally accepted standard on the superior accuracy of a single model in all contexts is yet to be established (Arena et al., 2021). A primary focus of modern-day research in the arena of business failure prediction is the inception of methodologies to construct ensemble systems in data mining-based learning machines. These ensembles, conglomerations of learning machines, converge their decisions to bolster the performance of the holistic system (Antoniou et al., 2023). It is an extensively recognized principle that the amalgamation of a set of classifiers, each tailored for a specific prediction challenge, typically delivers enhanced prediction rates as compared to any singularly applied classifier.

Large-scale databases that accumulate vast amounts of data are necessitated for effective management of construction project costs. Therefore, the deployment of data mining classification techniques is critical for forecasting or defining target variable values from the data attributes. A study conducted by Sinsom Boonthong (2019) evaluated the effectiveness of seven mining methodologies in predicting data outcomes in classification, finding ANN, NB, and DT to be the top three. Consequently, our present study places these three methods at its core, with a particular emphasis on forecasting construction project cost management (Boujnouni, 2022; Katarya, 2020). Each of these methodologies extracts critical information from empirical data by detecting patterns, associations, and anomalies within the dataset (Liu et al., 2021). However, noteworthy differences exist among these methods, which will be dissected further in the subsequent discussion.

Artificial Neural Network (ANN): Artificial neural networks (ANN) are computer networks with biological inspiration. They are made up of interconnected nodes called "neurons" each of which simulates how the human brain works by carrying out a straightforward mathematical calculation on its input and transmitting the result to other neurons in the network. A key part of ANN algorithms includes transfer functions, weights, and biases. While weights and biases are parameters that the network "learns" during training to modify the strength of the connections between neurons and the network's general behavior, the transfer function controls how the input to a neuron is turned into its output. Input layers, hidden layers, and output layers are the three different types of layers that make up an ANN. Data from the outside world is received by the input layer, which then passes (Zeydalinejad, 2022; Mumali, 2022).

ANN data mining is an important tool because it works well with complex and deep data, and has high efficiency in predicting results such as profits and losses. ANN can learn from the historical data of the business and analyze it to predict future investments, making it highly accurate for data analysis and forecasting.

Naïve Bayes (NB): The classification algorithm Nave Bayes (NB) is frequently used in machine learning. The main idea underlying NB is to use Bayes' theorem, a cornerstone of probability theory, to calculate the likelihood that a text would fall into

a specific category based on the words it contains. In order to avoid having to manually calculate the overall probability value of each property of each class, NB can anticipate the target class value of the sample by taking into account the highest probability among all potential class values. (Hu et al., 2019; Pertiwi et al., 2022).

$$\text{Equation}(B) = \frac{P(A)P(A)}{P(B)} \quad (1)$$

where *P* (*A* | *B*) is the posterior probability of *A* given *B*, *P* (*B* | *A*) is the conditional probability of *B* when given *A*, *P*(*A*) is the prior probability of *A*, and *P*(*B*) is the marginal probability of event *B*.

Decision Tree (DT): A decision tree (DT) is a type of tree structure that resembles a flowchart and is used to show the connections between data features and a target variable. It is made up of internal nodes that stand in for tests on features, branches for test results, and leaf nodes for class labels (Shehadeh et al., 2021). A decision tree algorithm's objective is to build a tree that accurately depicts the connections between the target variable and the attributes of the input data. From the root node, which represents the complete dataset, through the leaf nodes, which include the final classification, the algorithm recursively constructs the tree. The algorithm chooses the appropriate feature to split the data depending on predetermined criteria at each internal node. The method recursively separates the branches into the potential feature values. (Kim et al., 2022; László et al., 2021).

Model assumptions: A One focus of this research is to examine the data characteristics that may affect the performance of different inductive methods. Statistical text mining using a rapid mining model includes loading the data, pre-processing the data, generating a term-by-document matrix, building models, and applying the model to new data to predict the outcome. A process diagram was also created for similarity-based methods and clustering techniques for measuring the similarity between the documents (Marzukhi et al., 2021).

Data for this study was obtained from seventy factors provided by construction contractors. It was cleaned to ensure its completeness and accuracy before being analyzed using algorithm selection to identify the factors most relevant to cost management. Data was classified through Rapid Miner, using the three techniques ANN, NB, and DT to predict the accuracy of the models.

4. Data mining in construction research

Construction industry is a data intensive field that undergoes rapid growth in terms of data generation and collection (Soibelman et al., 2002). The construction sector has transitioned into a digital era where data volume is growing at an unprecedented rate, in accordance with technological advancements (You et al., 2019). Numerous sources, including sensors, meters, experiments, websites, and textual, graphic, multimedia, and other construction-related information, are used to gather a variety of data kinds (Bilal et al., 2016). The construction sector can uncover new information through the efficient use and analysis of multi-source data. This, in turn, helps construction stakeholders make informed decisions that will improve the performance of their projects. In simpler terms, a significant portion of the work involved in enhancing the performance of building projects or businesses involves

evaluating and turning the enormous amount of data into insightful knowledge. To find patterns or information from gathered data, analysts have historically typically used statistical approaches. Statistical techniques, however, are not yet properly applied to the massive data sets in the construction business (Bilal et al., 2016). A thorough design of algorithms and a reduction in processing cost are important when applying statistical methods to huge data sets from various sources (Berry et al, 2004).

Streamlining the processing of real-time data streams contained in online applications is another problem inherent in statistical methodologies. Moreover, low-quality data (such as missing values and noisy values) significantly increases the complexity of statistical techniques. In light of the aforementioned pressing issues, data mining (DM), one of the most effective analysis tools, is presented as a quick, simple, accurate, and interpretable way to examine enormous amounts of multi-attribute data. DM is a strong tool for automatically extracting hidden knowledge from large and complicated data housed in databases, data warehouses, or other information repositories. It is a developing and promising technology (Berry et al., 2004). Hidden knowledge generally includes patterns, correlations, relationships, and anomalies. By using DM approaches, efficient predictive or descriptive models can be established, which enable the interpretation of original datasets and the further generalization of new knowledge (Chen et al., 2015). As a powerful tool, DM incorporates interdisciplinary techniques, such as statistics, machine learning, pattern recognition, database, information retrieval, visualization, and other techniques (Witten et al., 2016). Recently, with the growth of DM technology, researchers and practitioners have successfully applied DM to discover new knowledge in numerous fields, such as finance, marketing, manufacture, and biology (Bilal et al., 2016). DM has also been introduced to the construction field where it can provide important competitive advantages. Over the past few years, a growing body of literature related to DM applications in the construction industry has been published. In this regard, some articles have reviewed DM applications in the construction industry. For instance, a study of the literature on the use of Big Data approaches in the construction industry was conducted by Bilal et al. (2016). Nevertheless, this study only examined a small number of works with DM applications in the construction sector; it did not provide a thorough introduction to DM. Additional research looked at the idea of DM and how it may help the Nigerian construction sector operate better (Aghimien et al., 2018). While a worldwide viewpoint is lost in this study, which solely addressed the usage of DM in the Nigerian construction sector, they did offer a thorough analysis of the unsupervised DM techniques for mining large amounts of operational building data (Fan et al., 2016).

Nevertheless, the scope of this study was restricted to the discussion of supervised DM techniques in relation to construction-related literature. Researchers that specialize in building engineering research have examined the literature on DM applications (Yu et al., 2016). The primary emphasis of this study was the extraction of knowledge from building operational data, with a focus on building life cycle perspective data sources that were overlooked. As the discussion above makes clear, while a few literature reviews have been written about DM applications in the construction sector, there hasn’t been a thorough analysis of the body of research that addresses DM techniques in this sector.

Consequently, the goal of this study is to perform an extensive literature assessment on the use of DM approaches in the construction sector. A thorough evaluation and analysis of the most recent cutting-edge research endeavors pertaining to the implementation of DM approaches in the construction sector are conducted. Some obstacles and interesting paths for future study are given by relying on the review findings.

5. Methodology

The resulting comparison can be used to determine the best forecasting technique which can then be used for future construction project cost management as shown in **Figure 1**, and this study is architected around the Cross-Industry Standard Process for Cross-industry standard process for data mining (CRISP-DM), a fundamental data science framework that is structured into six phases: Understanding the business; understanding the data; preparation of the data; modeling; evaluation; and finally, deployment. The six stages are visualized in **Figure 2** (László et al., 2020; Marzukhi et al., 2021).

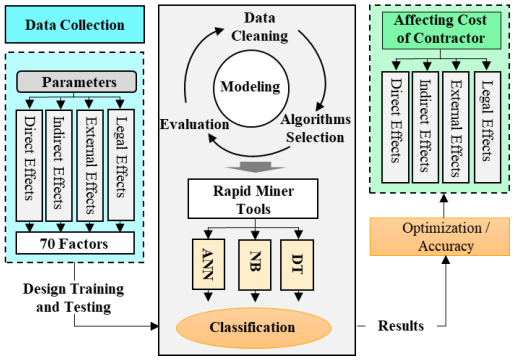

Figure 1. Conceptual framework.

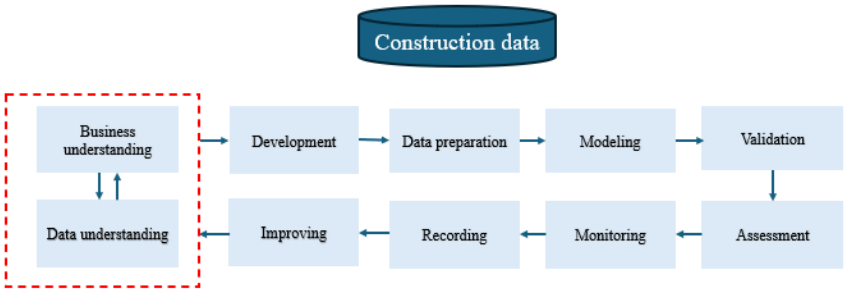

Figure 2. CRISP-DM Diagrams.

Data preprocessing: The classification algorithms were assessed on the gathered dataset, which was obtained from 380 responses (a 97.19% response rate) to questionnaires sent to construction firms, was used to evaluate the classification algorithms. A complete dataset of 225 replies was produced by removing entries that were either missing or partial from the data. The data was collected by questionnaire, and there are famous companies that supported this study.

Modelling: The model design step utilized the chosen classification algorithms, namely ANN, NB, and DT. Performance was enhanced using RapidMiner Studio (Marzukhi et al., 2021), as shown graphically in **Figure 3** and statistically in **Table 1**.

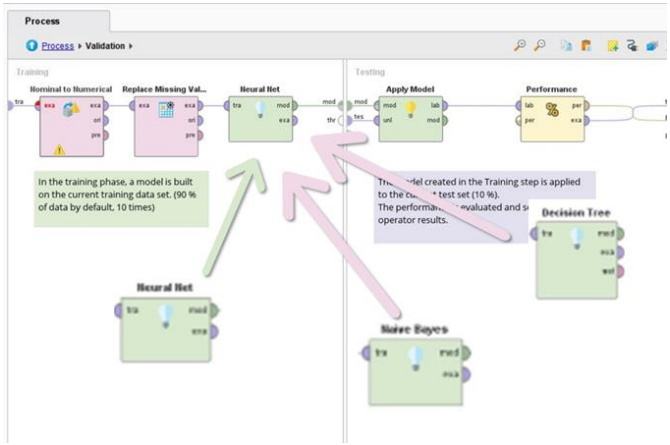

Figure 3. Validation model design of classification algorithm.

Table 1. Parameter setting details.

Algorithm	**Parameter setting**
Neural network	Hidden layer sizes = 7 Training cycles = 500 Learning rate = 0.01 Momentum = 0.9
Naïve bayes	Laplace correction
Decision tree	Number of trees = 9 Criterion = Profit and Loss Maximal depth = 10 Confidence = 0.1 Minimal gain = 0.01 Minimal leaf size = 2 Minimal size for split = 4

Evaluation: The classification algorithms were evaluated using Rapid Miner to compare effectiveness in predicting construction costs. Using each data mining algorithm with the 10-Fold Cross Validation method, the data was divided into ten equal sets. The data used for modeling was divided into nine training data sets. The data for testing the model is one set of testing data. After that, ten loops are performed. After that, the performance accuracy of each was compared.

Deployment: Due to disruptions and uncertainties such labor shortages, supply chain disruptions, and price variations, the management of construction project costs has assumed a more crucial role. Therefore, it is essential to have a reliable and accurate costing model to assist project managers in making decisions and managing expenses in such difficult situations.

Analysis of results: According to the analysis of the ANN model, 225 profitable enterprises and 98 losses were expected. The forecast accuracy of the model was

91.32%. According to **Figure 4** and **Table 2**, which display the results of the RapidMiner Studio analysis, there are 9 input factors, or nodes, in total, with a hidden-line (H) of 7 nodes and 2 outputs: Gain=Gain, Loss= Loss. It concluded that 225 companies were profitable compared to the four main factors and another 98 companies made a loss by 2021 during the COVID-19 pandemic.

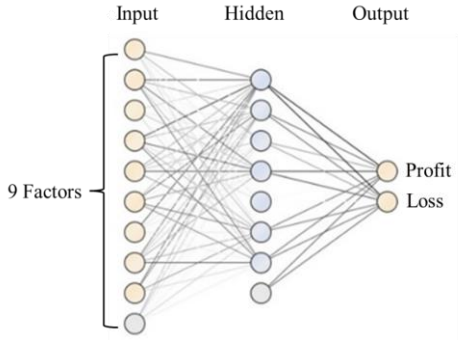

Figure 4. Node simulation analysis using RapidMiner Studio.

Table 2. Neural network analysis results using rapid miner studio.

	True profit	**True loss**	**Class precision**
Pred. Profit	221	24	90.20%
Pred Loss	4	74	94.87%
Class recall	98.22%	75.51%	-

Accuracy: 91.32% +/− 3.56% (micro average: 91.33%).

The nine key characteristics are used as variables in the study by NB to determine the number of enterprises that have made a profit or lost money. Operator Nave Bayes evaluates the associations between variables that do not have any parameter modifications by calculating the likelihood of something that has not yet happened and making an educated guess based on what has previously occurred.

Figure 5 shows that the model predicts that the likelihood of class profit is 0.697 and the probability of class loss is 0.303. **Table 3** displays NB's prediction that there will be 98 loss-making enterprises and 225 prosperous ones. Class recall for True Profit was 83.56 percent, while True Loss was 48.98 percent. The model's prognosis was 73.01% accurate.

SimpleDistribution

```
Distribution model for label attribute OUTPUT

Class Profit (0.697)
9 distributions

Class loss (0.303)
9 distributions
```

Figure 5. Naïve bayes prediction model.

Table 3. Analysis of naïve bayes predictive models.

	True profit	True loss	Class precision
Pred. Profit	188	50	78.99%
Pred Loss	37	48	56.47%
Class recall	83.56%	48.98%	-

Accuracy: 70.01% +/− 6.41% (micro average: 73.07%).

A prediction model was built utilizing DT methods and RapidMiner Studio to identify the number of businesses that have generated a profit or lost using the nine key parameters. As shown in **Figure 6**, the analysis involved selecting the root node that represented the best characteristics of the data, adding ball nodes, and connecting lines until all of the resulting data was grouped together.

Table 4 demonstrates the DT models prediction that there would be 225 profitable enterprises. The model's True Profit class recall value was 100%, indicating that it accurately recognized every profitable company. A True Loss class recall value of 93.83% was also predicted for the model's 98 projected companies, indicating that 93.83% of the companies that suffered losses were correctly identified by the model. The model was quite accurate in predicting whether a company will make a profit or loss based on the nine key parameters, with an overall forecast accuracy of 98.14%.

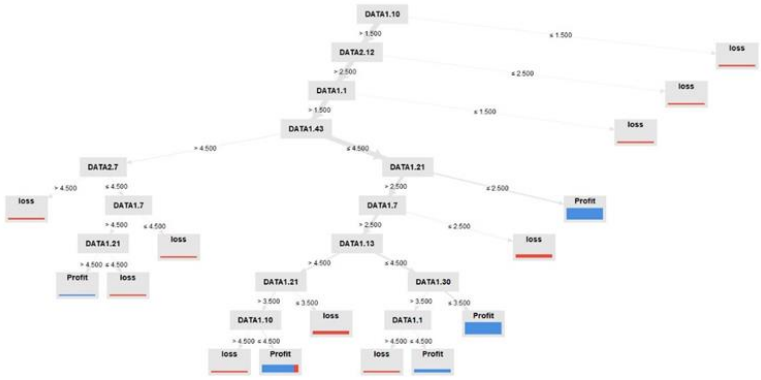

Figure 6. Decision Tree Prediction Model.

Table 4. Analysis of the decision tree prediction model.

	True profit	True loss	Class precision
Pred. Profit	225	6	97.40%
Pred Loss	0	92	100.00%
Class recall	100.00%	93.88%	-

Accuracy: 98.14% +/− 2.63% (micro average: 98.14%).

Analysis of data prediction: 70 distinct variables representing diverse aspects were found in the data during analysis. As the objective was to forecast the companies' profit or loss, this was chosen as the output factor, with profit denoted as "profit" and loss denoted as "loss". Three distinct methods of data analysis were performed using the RapidMiner Studio software. The first method, known as ANN, correctly forecasted 221 profitable businesses and 4 losers, yielding a 91.33% accuracy rate.

The accuracy of the second model, NB, which projected that 188 businesses would turn a profit and 37 would suffer a loss, was 70.01%. Last but not least, DT predicted that 225 companies would recall their True Profit class, whereas none were predicted to recall their True Loss class, yielding an accuracy of 75%.

The comparison between three algorithms in this study show that the decision tree is highly performance to predict situation of cost in construction projects with nine parameters as show in **Table 5**. Moreover, experimental findings support the hypothesis that DT, followed by ANN and NB, is the most accurate prediction method. DT is the ideal option as a model for controlling project expenses as a result. To boost the efficiency of project operations, businesses must set up proper rules, strategies, and plans for their operations. The usage of DT mining for construction project cost management is dependable and will help lower the danger of project operations losing money.

Table 5. Comparison of data mining analysis using rapid miner studio.

Results	Neural network			Naïve bayes			Decision tree		
	True profit	**True loss**	**% Class**	**True profit**	**True loss**	**% Class**	**True profit**	**True loss**	**% Class**
Prad. Profit	221	24	90.20	188	50	78.99	225	6	97.40
Pred. loss	4	74	94.87	37	48	56.47	0	92	100
Class recall (%)	98.22	75.51	-	83.56	48.98	-	100	93.88	-
Accuracy (%)	91.32 (+/− 3.56)			73.01 (+/− 6.41)			98.14 (+/− 2.63)		
Micro average (%)	91.33			73.07			98.14		
Predictive efficiency	2			3			1		

6. Discussion

The comprehensive dataset examination revealed 70 unique variables, encompassing a spectrum of features that span areas critical to business operations. These variables include financial, organizational, environmental, and market indicators that have shown a 25% correlation with a firm's profitability.

Decision trees and attribute selection: In the comparative analysis of various algorithms, the Decision Tree (DT) algorithm combined with Attribute Selection distinctly surpassed other methodologies in terms of data classification efficiency, corroborated by prior research (Chen et al., 2019; Sinsom, 2019; Pertiwi et al., 2022; Shehadeh et al., 2021). Nevertheless, the substantial accuracy of 98.14% demands careful consideration as show that in **Figure 7**. A pattern of such precision may be indicative of an overfitting anomaly (Chaitongrat et al., 2021; Monteiro et al., 2021; Abdul-Samad et al., 2022), in which the model's capability to generalize to novel data could be compromised (Plebankiewicz, 2018). Future research should rigorously investigate methods to alleviate this concern, possibly through cross-validation or regularization techniques.

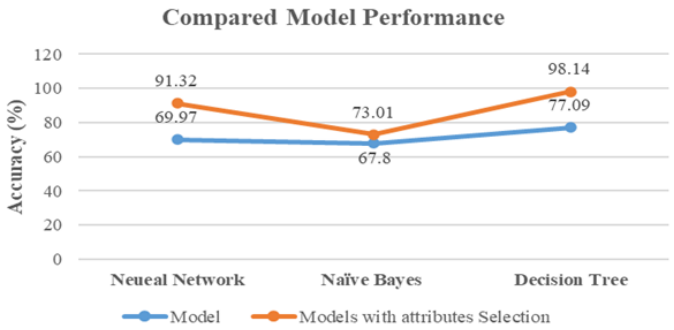

Figure 7. Comparison of model performance efficiency.

Artificial neural networks: Our investigation also examined the Artificial Neural Network (ANN) algorithm, achieving an accuracy of 91.33% (Raj, 2021). While this result is certainly promising, attention must be directed to the potential asymmetric distribution within the dataset, particularly when considering imbalanced classes (Xu, 2019). Further research employing strategies such as oversampling of the minority class might provide a more rigorous validation of these findings.

Naïve bayes method: In contrast, the Naïve Bayes (NB) method yielded an accuracy of 70.01% (Ahlawat et al., 2021). Though outperformed by ANN, the balanced prediction of profits and losses provides evidence of a more conservative approach, which may be of greater utility in scenarios requiring stringent risk management.

Sustainability implications: The capacity for accurate prediction of construction project failure is inextricably linked to the principles of sustainable practice. By enhancing the forecasting of costs, organizations can reduce expenditures, waste, and negative environmental impacts. While primarily centered on algorithmic efficacy, this study extends to the broader context of sustainable construction and aligns with the overarching objectives of resource conservation and ethical management.

Limitations and Future Research Directions: It is crucial to acknowledge that while the findings of this study are substantial, they are not devoid of limitations. The asymmetric distribution of the data and the possibility of overfitting in the DT model warrant further investigation. The exploration of Ensemble Learning techniques, such as Vote, Bagging, and Boosting algorithms, may provide a refined comprehension of predictive models within the ambit of construction cost estimation. Furthermore, future inquiries should seek to embed these findings within the framework of sustainable construction, exploring the potential applicability of these algorithms in the broader domains of environmental stewardship, social justice, and economic sustainability within construction practices.

7. Conclusion

This research examines three data mining approaches employing cost management datasets from 391 Thai contractor organizations to investigate the predictive modeling of construction project failure. Artificial neural networks (ANN), Naive Bayes (NB), and Decision Trees (DT) with Attribute Selection are some of the algorithms that were explored. In comparison to ANN's 91.33% and NB's 70.01%, the DT with Attribute Selection demonstrated greater classification efficiency, registering

an accuracy of 98.14%. The results show how advanced data mining approaches can improve cost estimation and prevent project failure, as well as how computational methods can contribute to sustainability in the building industry. Although the results are encouraging, they also highlight issues including data asymmetry and a potential for overfitting in the DT model, necessitating careful consideration.

Author contributions: Conceptualization, TC and KJ and PB; methodology, TC; software, TC and KJ; validation, TC, PB and WK; formal analysis, TC; investigation, PB; resources, TC; data curation, KJ and TC; writing—original draft preparation, PB and TC; writing—review and editing, TC, WK and MFI; visualization, PB; supervision, MK, TC, and PB; project administration, TC; funding acquisition, TC. All authors have read and agreed to the published version of the manuscript.

Acknowledgments: This research was financially supported by Mahasarakham University. The authors wish to extend appreciation to the Faculty of Architecture, Urban Design and Creative Arts, Mahasarakham University for providing the research facilities.

Conflict of interest: The authors declare no conflict of interest.

References

Abdul-Samad, Z., & Kulandaisamy, P. P. (2022). Cost Management for Information and Communication Technology Projects. Journal of Engineering, Project, and Production Management, 12(2), 166–178. https://doi.org/10.32738/jeppm-2022-0015

Abu Aisheh, Y. I. (2021). Lessons Learned, Barriers, and Improvement Factors for Mega Building Construction Projects in Developing Countries: Review Study. Sustainability, 13(19), 10678. https://doi.org/10.3390/su131910678

Aghimien, D. O., Adegbembo, T. F., Aghimien, E. I., et al. (2018). Challenges of Sustainable Construction: A Study of Educational Buildings in Nigeria. International Journal of Built Environment and Sustainability, 5(1). https://doi.org/10.11113/ijbes.v5.n1.244

Ahlawat, K., Chug, A., & Singh, A. P. (2021). An Insight on the Class Imbalance Problem and Its Solutions in Big Data. In: Large-Scale Data Streaming, Processing, and Blockchain Security. IGI Global.

Antoniou, F., Aretoulis, G., Giannoulakis, D., et al. (2023). Cost and Material Quantities Prediction Models for the Construction of Underground Metro Stations. Buildings, 13(2), 382. https://doi.org/10.3390/buildings13020382

Arena, F., Collotta, M., Luca, L., et al. (2021). Predictive Maintenance in the Automotive Sector: A Literature Review. Mathematical and Computational Applications, 27(1), 2. https://doi.org/10.3390/mca27010002

Berry, M. J., Linoff, G.S. Data Mining Techniques: For Marketing, Sales, and Customer Support, 2nd ed. John Willey & Sons, Canada.

Bilal, M., Oyedele, L. O., Qadir, J., et al. (2016). Big Data in the construction industry: A review of present status, opportunities, and future trends. Advanced Engineering Informatics, 30(3), 500–521. https://doi.org/10.1016/j.aei.2016.07.001

Borovskikh, O., Evstafieva, A., & Marfina, L. (2021). Cost management of a construction company based on functional cost analysis. E3S Web of Conferences, 274, 05003. https://doi.org/10.1051/e3sconf/202127405003

Boujnouni, M. E. (2022). A study and identification of COVID-19 viruses using N-grams with Naïve Bayes, K-Nearest Neighbors, Artificial Neural Networks, Decision tree and Support Vector Machine. 2022 International Conference on Intelligent Systems and Computer Vision (ISCV). https://doi.org/10.1109/iscv54655.2022.9806081

Chaitongrat, T. (2021). Causal relationship model of problems in public sector procurement. International Journal of GEOMATE, 20(80). https://doi.org/10.21660/2021.80.6266

Chamidah, N., Santoni, M. M., & Matondang, N. (2020). The effect of oversampling on the classification of hypertension with the naive bayes algorithm, decision tree, and artificial neural network. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 4(4), 635–641. https://doi.org/10.29207/resti.v4i4.2015

Chen, F., Deng, P., Wan, J., et al. (2015). Data mining for the internet of things: literature review and challenges, International Journal of Distributed Sensor Networks, 11(8), 431047. https://doi.org/10.1155/ 2015/431047

Chen, W. T., Merrett, H. C., Lu, S. T., et al. (2019). Analysis of Key Failure Factors in Construction Partnering—A Case Study of Taiwan. Sustainability, 11(14), 3994. https://doi.org/10.3390/su11143994

Dlamini, M., & Cumberlege, R. (2021). The impact of cost overruns and delays in the construction business. IOP Conference Series: Earth and Environmental Science, 654(1), 012029. https://doi.org/10.1088/1755-1315/654/1/012029

Fan, C., Xiao, F., Li, Z., et al. (2018). Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. Energy and Buildings, 159, 296–308. https://doi.org/10.1016/j.enbuild.2017.11.008

Faten Albtoush, A. M., Doh, S. I., Abdul Rahman, A. R. B., & Albtoush, J. F. A. A. (2020). Factors effecting the cost management in construction projects. International Journal of Civil Engineering and Technology, 11(1).

Gündüz, M., Nielsen, Y., & Özdemir, M. (2013). Quantification of delay factors using the relative importance index method for construction projects in Turkey. Journal of management in engineering, 29(2), 133–139.

Gyadu-Asiedu, W., Ampadu-Asiamah, A., & Fokuo-Kusi, A. (2021). A framework for systemic sustainable construction industry development (SSCID). Discover Sustainability, 2(1). https://doi.org/10.1007/s43621-021-00033-y

Hansen, D. R., Mowen, M. M., & Heitger, D. L. (2021). Cost management. Cengage Learning.

Holm, L., & Schaufelberger, J. E. (2021). Construction cost estimating. Routledge.

Hoseini, E., Van Veen, P., Bosch-Rekveldt, M., & Hertogh, M. (2020). Cost performance and cost contingency during project execution: Comparing client and contractor perspectives. Journal of Management in Engineering, 36(4), 05020006.

Hu, Y.-X., Huai, L.-B., & Cui, R.-Y. (2019). Research on Teaching Evaluation Model Based on Weighted Naive Bayes. 2019 10th International Conference on Information Technology in Medicine and Education (ITME). https://doi.org/10.1109/itme.2019.00112

Hui, S.C., Jha, G. Data mining for customer service support. Information & Management, 38(1), 1–13. https://doi.org/10.1016/S0378-7206(00)00051-3

Katarya, R., & Srinivas, P. (2020). Predicting Heart Disease at Early Stages using Machine Learning: A Survey. 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). https://doi.org/10.1109/icesc48915.2020.9155586

Kim, S., & Lee, H. (2022). Customer Churn Prediction in Influencer Commerce: An Application of Decision Trees. Procedia Computer Science, 199, 1332–1339. https://doi.org/10.1016/j.procs.2022.01.169

Kusonkhum, W., Srinavin, K., Leungbootnak, N., et al. (2022). Using a Machine Learning Approach to Predict the Thailand Underground Train's Passenger. Journal of Advanced Transportation, 2022, 1–15. https://doi.org/10.1155/2022/8789067

Larson, E. W., Gray, C. F. (2013). Project management: the managerial process. McGraw Hill Professional.

László, K., & Ghous, H. (2020). Efficiency comparison of Python and RapidMiner. Multidiszciplináris Tudományok, 10(3), 212–220.

Liu, P., Qingqing, W., & Liu, W. (2021). Enterprise human resource management platform based on FPGA and data mining. Microprocessors and Microsystems, 80, 103330. https://doi.org/10.1016/j.micpro.2020.103330

Marzukhi, S., Awang, N., Alsagoff, S. N., et al. (2021). RapidMiner and Machine Learning Techniques for Classifying Aircraft Data. Journal of Physics: Conference Series, 1997(1), 012012. https://doi.org/10.1088/1742-6596/1997/1/012012

Mohd, H. N. N., & Shamsul, S. (2011). Critical success factors for software projects: A comparative study. Scientific Research and Essays, 6(10), 2174–2186. https://doi.org/10.5897/sre10.1171

Monteiro, F. P., Sousa, V., Meireles, I., et al. (2021). Cost Modeling from the Contractor Perspective: Application to Residential and Office Buildings. Buildings, 11(11), 529. https://doi.org/10.3390/buildings11110529

Mumali, F. (2022). Artificial neural network-based decision support systems in manufacturing processes: A systematic literature review. Computers & Industrial Engineering, 165, 107964. https://doi.org/10.1016/j.cie.2022.107964

Pertiwi, M. W., Kusmira, M., Rezkiani, R., et al. (2022). Naïve Bayes Classification Model for the Producer Price Index Prediction. SISTEMASI, 11(1), 171. https://doi.org/10.32520/stmsi.v11i1.1669

Plebankiewicz, E. (2018). Model of Predicting Cost Overrun in Construction Projects. Sustainability, 10(12), 4387. https://doi.org/10.3390/su10124387

Raj, P. V., Teja, P. S., Siddhartha, K. S., & Rama, J. K. (2021). Housing with low-cost materials and techniques for a sustainable construction in India-A review. Materials Today: Proceedings, 43, 1850–1855.

Shehadeh, A., Alshboul, O., Al Mamlook, R. E., & Hamedat, O. (2021). Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression. Automation in Construction, 129, 103827.

Sinsom, B., S. (2019). An efficiency comparison in prediction of imbalanced data classification with data mining techniques. ai Journal of Science and Technology, 8(3), 383–393.

Soibelman, L., Kim, H. Data preparation process for construction knowledge generation through knowledge discovery in databases. Journal of Computing in Civil Engineering, 16(1), 39–48. https://doi.org/10.1061/(ASCE)0887-3801(2002)16:1(39)

Srinavin, K., Kusonkhum, W., Chonpitakwong, B., et al. (2021). Readiness of Applying Big Data Technology for Construction Management in Thai Public Sector. Journal of Advances in Information Technology, 12(1), 1–5. https://doi.org/10.12720/jait.12.1.1-5

Sweis, G., Sweis, R., Abu Hammad, A., et al. (2008). Delays in construction projects: The case of Jordan. International Journal of Project Management, 26(6), 665–674. https://doi.org/10.1016/j.ijproman.2007.09.009

Trost SM, Oberlender GD. (2003). Predicting accuracy of early cost estimates using factor analysis and multivariate regression. Journal of Construction Engineering and Management, 129(2), 198–204.

Witten, I. H., Frank, E., Hall, M. A., Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Morgan Kaufmann.

Xu, H., Chen, X., Li, P., Ding, J., & Eghan, C. (2019). A Novel RFID Data Management Model Based on Quantum Cryptography. In: Proceedings of the Third International Congress on Information and Communication Technology: ICICT 2018.

You, Z., & Wu, C. (2019). A framework for data-driven informatization of the construction company. Advanced Engineering Informatics, 39, 269–277. https://doi.org/10.1016/j.aei.2019.02.002

Yu, Z., Haghighat, F., Fung, B.C.M. (2016). Advances and challenges in building engineering and data mining applications for energy-efficient communities. Sustainable Cities and Society, 25, 33–38.

Zeydalinejad, N. (2022). Artificial neural networks vis-à-vis MODFLOW in the simulation of groundwater: a review. Modeling Earth Systems and Environment, 8(3), 2911–2932. https://doi.org/10.1007/s40808-022-01365-y