BioGecko

Vol 12 Issue 01 2023

ISNN NO: 2230 5807

A CASE STUDY ON PROJECT ANALYTICS RELATED TO DATA DRIVEN FRAME WORK BY USING MACHINE LEARNING

1. *DR.D.ANITHA KUMARI, Associate Professor, Department of CSM, TKRCET, Email id: anithakumaridara@gmail.com, Orchid id: 0000-0001-5506-7761*
2. *VENKATA CHARAN KANTUMUCHU, Global Quality Director, Electrex Inc, Dept. of Manufacturing Engineering, Bradley University, Peoria, IL, USA, vkantumuchu@mail.bradley.edu, https://orcid.org/0000-0002-54-6107*
3. *Dr G LAXMAIAH, Professor, Mechanical Engineering Department, Chaitanya Bharathi Institute of Technology, Hyderabad, India, E-Mail:glaxmaiah_mech@cbit.ac.in, https://orcid.org/0000-0001-9224-6232*
4. *K.SRIDHAR, DEPARTMENT OF MECHANICAL ENGINEERING, LENDI INSTITUTE OF ENGINEERING AND TECHNOLOGY, Vizianagaram, Andhra Pradesh*

Abstract:

Project analytics refers to the analytical processes used to make project delivery easier. The current methods emphasise looking back at data and figuring out the underlying connections so you may make more intelligent choices in the future. Despite the widespread use of machine learning algorithms to solve issues in many fields (e.g., improving the efficiency of construction project design), only some studies have examined current machine learning approaches in the construction industry's project delivery. So, this study aims to evaluate a particular collection of machine learning algorithms to further contribute to this convergence between artificial intelligence and the execution building project. To tackle issues in project analytics, this research provides a machine learning–based, data–driven research approach. As a follow-up, it gives a case study demonstrating how this paradigm might be used. In this example, different machine learning models (Python's Scikit-learn package) were tested and assessed using preexisting data from an open-source data repository on building projects and the frequencies of cost overruns. Project cost overrun frequency was the dependent variable, while the other 44 variables (ranging from materials to labour and contracts) were classified for processing by several machine

learning models. Models such as the support vector machine, logistic regression, k-nearest neighbour, random forest, stacking (ensemble), and artificial neural network are included. The best possible prediction model was found using various feature selection and assessment strategies, such as the Univariate feature selection, Recursive feature elimination, Select FromModel, and confusion matrix. It is also discussed in this study how the suggested research framework might be applied to various research settings in the field of project management. Practitioners, stakeholders, and academics would benefit significantly from the proposed framework, its illustrative example in the context of building projects, and its potential for adoption in many situations.

Keywords: project analytics, SVM, logistic regression, KNN

1. Introduction:

There can only be successful initiatives with the correct data and tools. Through analytics, project managers can make better choices at every project development stage. Earned value analysis and Monte

Carlo simulation are two examples of the statistical models used in project analytics. They play an important role in risk management and the actual project carrying. The use of project analytics is on the rise because of its many apparent benefits, such as the enhancement of foresight and prediction, comparison with similar projects, and identification of time-dependent 3-5 patterns. Recently, there has been a rise in curiosity about project analytics and how they may be applied to and facilitated by modern technological tools. There are five levels at which project analytics may be comprehended broadly. The first kind, descriptive analytics, involves looking back at past data. The second approach, diagnostic analytics, seeks to identify and clarify underlying causes and effects. The third kind of analytics actively strives to anticipate the future. The next stage is prescriptive analytics, which suggests the following actions after making a forecast. Finally, the purpose of cognitive analytics is to foresee potential issues. The first three tiers are the most amenable to technological implementation. Often, the information needed for the fourth and fifth phases is harder to get by because it is either less readily available or poorly structured.

While defining KPIs for a project might be difficult, it's made more accessible by looking for patterns in the data. Due to its immediate benefits to the primary baseline indicators focusing on productivity, profitability, cost, and time, it is projected that project analytics will continue to undergo progress. The field of project management is very fluid and adaptable, making it an ideal setting for applying machine learning algorithms.

Cognitive analytics, which includes machine learning in project analytics, focuses on anticipating issues. Machine learning studies how computers may learn to enhance their performance via exposure to previous examples or by acquiring new skills. It has the potential to improve upon the management community's current skills and methods for completing difficult tasks. New and improved machine learning algorithms and methods have been developed and introduced in recent years due to the field's widespread relevance and practical use. Computer vision, voice recognition, natural language processing, and robot control are examples of software that may be made possible by artificial intelligence. In the construction business, it is common to employ VR with BIM replication or risk prediction to keep tabs on work sites. Machine learning is also being used in other sectors, including consumer services (where it is increasing customer pleasure) and transportation (where it is decreasing human error). Classification, regression, ranking, clustering, dimensionality reduction, and manifold learning are recent applications and developments in machine learning. Linear predictors, boosting, stochastic gradient descent, kernel techniques, and closest neighbour are only some of the modern learning models available. To make education more accessible and efficient, developers are releasing new apps and learning methods.

Machine learning-based framework for project analytics:

Predictions of a categorical dependent variable are an everyday use of machine learning models in scientific inquiry. Therefore, if the objective variable being measured is categorical, it may be used in project analytics. First, the objective variable has to be transformed into a definite form if it is not already one. If the project's cost is the aim or objective variable, we can make it a categorical variable by restricting it to one of two categories. To depict a cheap project, the first value may be set to 0, while a high-priced one could be represented by the second. It is possible to divide the total cost of the projects into two groups, one with low costs and one with high costs, based on the average or median cost value for the whole set of projects.

Machine learning models have their uses for data-driven decision making. This is because conventional statistical approaches (such ordinary least square (OLS) regression) need hypothesis testing before arriving at concrete equations for the intended objective outcome variables. Machine learning algorithms, in contrast to conventional statistical techniques, discover patterns in data on their own. For instance, because an OLS regression model requires the underlying data to be linear, it is not the best option for a non-linear but separable dataset. However, a machine learning model only has to examine the dataset once to get to the underlying classes. Machine learning models outperform conventional statistical approaches, as seen in Figure 1.

Similarly, if the underlying study dataset has several features or independent assessments, machine learning models become more persuasive. Regardless of the features' distributions or collinearity, such models may isolate the most important ones for the associated classification performance. When there is a connection between independent variables, the findings from conventional statistical approaches might be skewed. There hasn't been a lot of research done recently using machine learning that focuses on project analytics. Nevertheless, there has been peripheral research on using AI to enhance cost estimates and risk prediction. Existing processes have also been optimised using models .

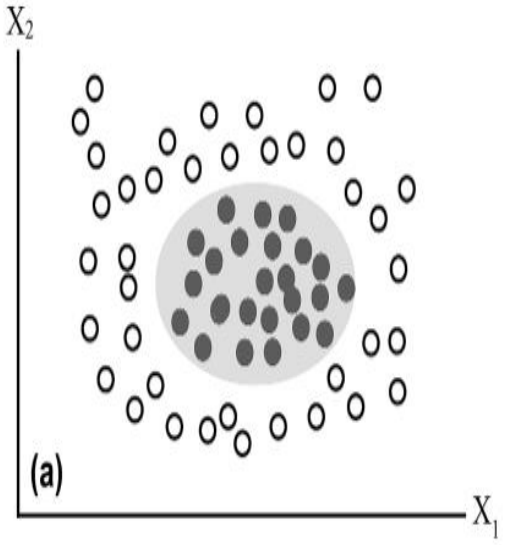

Figure 1: This example uses a generic dataset with two features to highlight the advantages of machine learning over conventional statistical modelling (X1 and X2). The abstract data set consists of two types of points, one of which is shown by a clear circle and the other by a black one. These values are not linear, but they may be grouped together. Normal statistical methods (such conventional least squares regression) will not be able to properly classify these values. However, any machine learning model can reliably distinguish between them.

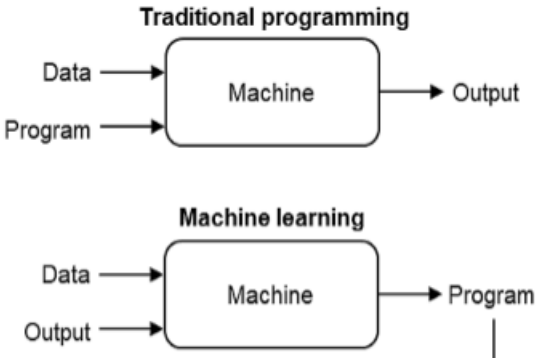

Figure 2: Traditional programming versus machine learning

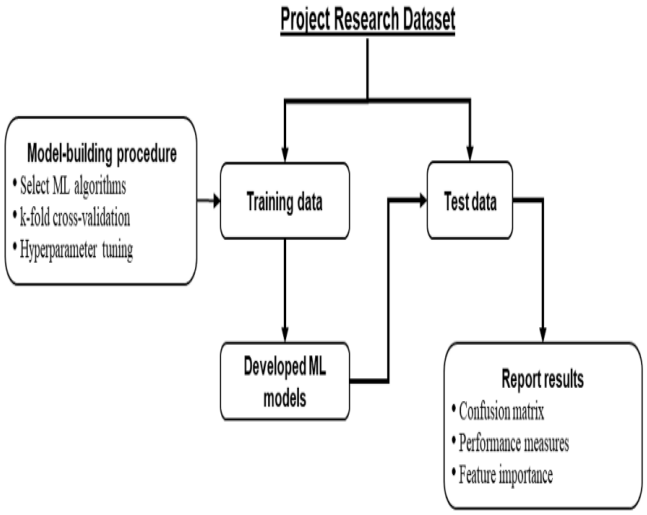

Figure 3. The proposed machine learning-based data-driven framework.

In order to get the desired result. Machine learning is not like conventional programming in that it uses an algorithm to generate a programme based on input data and its related output. As a consequence, the resulting software may be utilised to get valuable insights into the data pattern and make predictions. Machine learning is shown in Figure 2 in contrast to more conventional programming techniques.

in order to get a result. Machine learning is not like conventional programming in that it uses an algorithm to generate a programme based on input data and its related output. As a consequence, the resulting software may be utilised to get valuable insights into the data pattern and make predictions. Figure 2(b) contrasts machine learning with more conventional programming techniques.

Structure suggested using machine learning. This study's suggested research framework is shown in Figure 3 and is based on machine learning. The methodology starts off with the separation of the project research dataset into a training and test set. The study dataset may include several nominal and/or categorical independent variables, as was described above, but the research's sole dependent

variable must be categorical. Although there is no hard and fast rule for this split, the training data amount is often more than or equal to 50% of the original dataset.

Variables with solely numeric outputs are manageable by machine learning techniques. Therefore, we must first transform them into the correct numeric values when one or more of the underlying category variables have a textual or string result. Let's pretend there are only three possible values for a variable in terms of text: low, medium, and high. A simple example would be to use the number 1 to indicate a low value, 2 for a moderate value, and 3 for a high value. The RIDIT (relative to an identified distribution) score is only one example of a statistical method that may be used to transform ranked categories into numeric values. Parametrically, RIDIT compares the probabilities of two sets to find statistical differences between ranked categories. The subsequent sections of this paper provide a high-level overview of the other parts of the proposed framework.

Building a model is a standard practise. The framework then directs us to the model-building phase, where we use the gathered training data to create the ML models we want. The approach begins with the selection of appropriate machine learning algorithms or models. A few examples of popular machine learning algorithms include the support vector machine, logistic regression, k-nearest neighbours, artificial neural network, decision tree, and random forest. If an ensemble machine learning model is what you're after, you may use it instead. Better prediction performance than could be gained from any of the component learning models alone is the goal of an ensemble machine learning approach, which employs many algorithms or the same algorithm numerous times. As for ensemble methods, bagging, boosting, and stacking are three of the most used. Bagging is the process of splitting a large dataset into smaller, uniform samples for analysis. These selections are subsequently put through the machine learning algorithm's categorization process. To address the shortcomings of the currently used model, boosting selects a random subset of the dataset for fitting and training using a series of alternative models in sequence. Stacking heterogeneously combined many low-quality machine learning models to boost their prediction abilities.

The random forest technique, for instance, is a combination of many distinct kinds of decision tree models.

The k-fold cross-validation method will next be used to each of the machine learning models that passed the first step. K-fold cross-validation is a method where the training data is split up into k different groups. The (k1) folds in an iteration are used to train the chosen machine models, while the remaining (k1) fold is utilised for validation. Each of the k folds will be validated once throughout this procedure, which will be repeated until all k folds have been utilised. Results from these rounds are averaged to determine the prediction efficacy of the trained models. The standard deviation of the outcomes over several iterations is also used as a measure of predictive training efficiency, supplementing the average value. The k-fold cross-validation is shown in Supplemental Fig. 1.

Hyperparameter tuning, the third point, is necessary for most machine learning algorithms due to the need for a fixed value for each of their parameters. The performance of the underlying algorithm is very sensitive to the values chosen for these parameters. Each dataset may have a unique optimum value for these parameters when using a certain machine learning method. Repeatedly running the same method with varying values for its parameters is required in order to determine what that value

should be for a specific dataset. Grid search is only one of several techniques described in the literature for determining the best setting for a variable. Grid search works by segmenting hyperparameters into these grids. Various permutations of the model's inputs are represented by each grid point. In other words, the optimum parameter values are the ones that lead to the best performance.

Conducting tests on the created models and documenting the outcomes. After developing the necessary machine learning models from the training data, they must be put to the test on the test data. Afterwards, the underlying trained model is used to make predictions about the dependent variable for each data occurrence. As a result, for every data instance, in addition to the actual category for its dependent variable, we will also get the category as predicted by the underlying trained model. The machine learning model's performance is summarised by comparing the anticipated and actual values for the corresponding category outcomes.

The confusion matrix, which takes on four integer values, is the primary instrument for reporting outcomes from machine learning models. For the first metric, we're interested in the proportion of positive examples that the underlying trained model properly classified as positive (true-positive). The second figure represents the total amount of false-positive results (false-negative). Number of false-positive results (cases classified as negative) (false-positive). The fourth figure represents the proportion of false-positives that were recognised as such (true-negative). Scientists also publish machine learning outcomes using a few performance indicators based on the confusion matrix's four values. Accuracy, defined as the ratio of true predictions (positive and negative) to the total number of data instances (four-value sum of the confusion matrix), is the most used metric. Precision, recall, and F1-score are also often employed when reporting machine learning outcomes. It is common practise to utilise the precision of a model to indicate the quality of a positive prediction by calculating the ratio of true-positives to the entire number of positive predictions (i.e., true-positive + false-positive). By dividing the number of positive predictions by the total number of data instances that should have been predicted as positive (i.e., true-positive + false-negative), we get recall, also known as the true-positive rate.

The error rate equals F1-score, which is the harmonic mean of the previous two measurements, Precision and Recall (1-Accuracy).

Variable or feature significance is another crucial technique for presenting machine learning findings, since it identifies a set of independent factors (features) that most significantly affect classification performance. The significance of a variable is measured by how much that variable is relied upon by a particular machine learning algorithm to provide reliable predictions. Principal component analysis is the standard method for determining which variables are most important. It decreases the data's dimensionality while keeping information loss to a minimum, leading to a more transparent machine learning result. In addition, it is useful for visualising 2D and 3D data and identifying the most salient characteristics in a dataset.

Case study:

Use of the suggested framework in a specific case In this example, we show how the framework suggested in Fig. 2 might be used in the context of a building project. With this structure in place, we may divide tasks into two categories according to their propensity for cost overruns.

A delay of the first kind is relatively uncommon in projects (Rare class). The second kind represents perennially late initiatives (Often class). To achieve this, we take into account a number of outside factors.

Input data. The data used in this study was obtained from Kaggle, a public database. Including 44 independent factors or characteristics and 1 dependent variable, this survey-based dataset was compiled to investigate the root reasons of project cost overrun in Indian construction projects45. Factors like as materials and labour, contractual difficulties, and the scope of work are just few of the many examples of independent variables that may cause cost overruns.

Frequency of project cost overruns is the dependent variable (rare or often). There are a total of 139 observations in the dataset, 65 of which are uncommon and the rest 74 frequent. As part of getting the dataset ready for machine learning analysis, we transformed all of the categorical variables that may have had a textual or string result into a suitable numerical value range. We utilised the numbers 1 and 2 to signify the seldom and frequent categories, respectively. Fig. 2 of the supplementary material displays the correlation matrix between the 44 characteristics.

Machine learning algorithms:

Using the aforementioned research dataset, this study investigated four machine learning techniques to investigate the root reasons of project cost overruns. Here are several examples: support vector machine, logistic regression, k-nearest neighbours, and random forest.

A technique used to get insight from data is called a support vector machine (SVM). For instance, SVM would give a useful method for prediction if one wanted to determine and understand which projects are categorised as programmatically successful by means of the analysis of precedence data information. SVM operates by tagging objects with labels 56. By optimising marginal distances and minimising classification errors, these items are clustered into distinct classes based on their comparative features. Multi-dimensional plotting of the characteristics allows a separation line (a hyperplane; see supplemental Fig 3(a)) to be drawn, which serves to demarcate fundamental categories. The data points that are most centrally located in relation to the decision boundary on both sides of the debate are known as support vectors. They are the circles (both transparent and dark ones) in the vicinity of the hyperplane in Supplemental Fig 3(a). The hyperplane's location and orientation are critically determined by the support vectors.

To facilitate this procedure, a number of computational techniques are used, one of which is a kernel function to generate more derived characteristics. Although first developed for use with binary data, support vector machines may be expanded to handle more classes. In order to do this, individual SVMs are trained.

Logistic regression (LR) extends the capabilities of the linear regression model to foretell the presence or absence of a dichotomous variable, such as an event. See Supplemental Fig. 3(b) for an example of how a scatterplot is used to examine the association between a single independent variable and many dependent variables. In place of a straight line, the LR model sigmoidalizes the data. When crafting the model, the natural logarithm is taken into account. It returns a number between zero and one, which may be understood as the likeliness of belonging to the class in question. If you start with a bunch of rough estimations and refine them until you get a stable number, that's probably your best bet. When it comes to identifying and studying connections, LR is often the most basic option. When compared to regular regressions, it is more time and effort saving.

To find the most probable outcome, the k-nearest neighbours (KNN) method shows historical data and applies a fixed sample size (k). Using a distance metric, this strategy locates similar training samples. After considering all possible outcomes, a final categorization is formed based on the number of votes for the most popular option. There are three grey squares and one white square immediately next to the little circle, as shown in Supplemental Fig 3. Grey people make up the vast middle class. So, KNN will say that the instance (i.e., X) is grey. In contrast, the bigger circle in the same image has 10 white squares and four grey squares as its closest neighbours. White people make up a significant proportion of the upper class. So, the instance will be labelled as "white" by KNN. The benefits of KNN are that it may simplify the result and deal with missing data. In conclusion, KNN builds models based on similarities (and differences) and distances.

Multiple decision trees make up the machine learning process known as a random forest (RF). Decision trees are tree-like structures in which each node reflects a possible outcome of an evaluation of the input attribute. The interior nodes may be nested, with the leaf nodes representing the final decisions. It generates a classification result for a segment of the input vector that may be considered independent of the rest of the vector. It takes into account the mean value for continuous processes and the total number of votes in discrete processes52. Three decision trees are shown in Supplemental Fig 3 to demonstrate how a random forest works. Class B, Class A, and Class A are the results of Trees 1, 2, and 3, respectively. The ultimate verdict will be a class A as decided by the majority vote. Due to its focus on a subset of features, it may give unequal weight to certain characteristics because of this. In spite of its sensitivity to sample design, the random forest has the capacity to deal with multidimensionality and multicollinearity in data.

The function of the human brain is mimicked by artificial neural networks (ANN). Modeling logical propositions with weighted inputs, a transfer, and a single output (Supplementary Fig. 3) achieves this goal. In addition to its usefulness in modelling non-linear interactions, it also excels at managing multivariate data 62. There are primarily three channels via which ANN acquires knowledge. Among them are the supervised error-back propagation, the unsupervised Kohonen, and the supervised counter-propagation ANN. Both supervised and unsupervised ANNs exist. From the pharmaceutical industry to the electronics industry , ANN has found utility in a wide variety of settings. In addition to being very resilient to failure, it also picks up new skills quickly by observing others and organising itself .

To create the best possible model, machine learning ensemble approaches aggregate the results of many different basic classifiers. The goal of an ensemble approach is to increase model performance by taking into account several models and combining them into a single one. This single model will be stronger than the sum of its parts since it will have eliminated the shortcomings of each individual learner. There are two layers of classifiers in the stacking model, the basic classifier and the meta-learner. The training dataset is used to teach the basic classifiers, and a fresh dataset is created specifically for the meta-learner. Subsequently, the meta-classifier is trained with this updated data set. As shown in Supplemental Fig. 3, four models (SVM, LR, KNN, and RF) are used as base classifiers, and LR is used as a meta learner in this research.

(a) Training phase (values are in %)					
Machine learning algorithm	Training accuracy (standard deviation)				
Support vector machine	69.89 (9.09)				
Logistic regression	68.26 (9.39)				
k-nearest neighbours	76.98 (8.27)				
Random forest	78.14 (8.92)				
Stacking (ensemble) model	74.05 (9.56)				
Artificial neural network	67.50 (3.54)				
(b) Testing phase (values are in %)					
Machine learning algorithm	Accuracy	Precision	Recall	F1-Score	Error-rate
Support vector machine	72.50	65.00	76.47	70.27	27.50
Logistic regression	67.50	60.00	70.59	64.86	32.50
k-nearest neighbours	72.50	65.00	76.47	70.27	27.50
Random forest	77.50	68.18	88.24	76.92	22.50
Stacking (ensemble) model	70.00	63.16	70.59	66.67	30.00
Artificial neural network	72.50	65.00	76.47	70.27	27.50

Table 1. The performance of the six machine learning algorithms for the case study.

Feature optimisation approach	Accuracy	Precision	Recall	F1-score	Error-rate
Random forest with features from UFS	77.50	66.67	94.12	78.05	22.50
Random forest with features from RFF	72.50	63.64	82.35	71.19	27.50
Random forest with features from SFM	85.00	76.19	94.12	84.21	15.00

Table 2. The performance of the random forest algorithm from the testing phase using three different attribute/feature optimisation approaches. Values are in percentage.

Choosen characteristics. Feature selection refers to the process of identifying the most relevant features to use in order to improve model performance and reduce execution time. Three distinct

methods for selecting features are examined here. Univariate feature selection (UFS), recursive feature elimination (RFE), and the SelectFromModel (SFM) method. Each feature's significance in predicting the response variable 68 is evaluated independently by UFS. This approach is easy to learn and use, and it facilitates a more thorough comprehension of data. The chi-square values between characteristics are computed in this research. By fitting the model with all features in the dataset first, RFE is a sort of backwards feature removal in which the least important features are removed in turn. The model is then fit again until the parameter-specified number of features is retained. Using the relevance of features in the top-performing model, SFM selects the most useful features. The model's predictions on the training set serve as the basis for the threshold used in this method to pick features. Those traits are retained whose feature significance is higher than the threshold, while those with a lower feature value are omitted. After evaluating four different machine learning strategies, we use SFM in this research to further improve our results. After that, we re-train the top model using SFM characteristics.

Conclusions drawn from the case study. For the four chosen machine learning algorithms, we divided the dataset in half and used 70% for training and 30% for testing. To put these methods 70 into practise, we relied on Python's Scikit-learn module. We initially built six models using these six techniques and the training data. Specifically, we employed a five-fold validation and a goal to raise the accuracy value. The models were then used on test data. All necessary hyperparameter tunings for each algorithm were also performed to provide the highest quality classification results as feasible.

The results of the training and testing of each algorithm are summarised in Table 1. Supplementary Table 1 details the hyperparameters used by each method.

According to the results shown in Table 1, random forest achieved higher accuracy rates throughout the training and testing stages than the other three algorithms. For the training and testing stages, its accuracy was 78. and 77.5 percent, respectively. In the training phase, k-nearest neighbours performs second best (76.98%), while in the test phase, support vector machines, k-nearest neighbours, and artificial neural networks perform best (72.50-73.0%).

Since random forest performed the best, we centred our further investigations on this method. We used UFS, RFE, and SFM to optimise random forest features. Table 2 shows the final outcome. Among these three methods, SFM produces the most favourable results. Compared to USF and RFE, which have respective accuracies of 77.50% and 72.50%, its accuracy is 85.00%. Table 2 demonstrates that the SFM feature optimisation improves the testing phase accuracy from 77.50% in Table 1(b) to 85.00%. Data from Table 3:

Order	Feature
1	Delay in delivering material
2	Prices fluctuation
3	Shortage of labourers
4	Unavailability of equipment
5	Construction cost underestimation
6	Delayed payment
7	Cash flow problem
8	High rate of interest
9	Increase in salaries
10	Change design
11	Errors and omissions in design
12	Inaccurate quantity take-off
13	Delays in issuing information
14	Delays in decisions making
15	Insufficient time for documents
16	Extension of time
17	Rework due to error in the execution
18	Accidents during construction
19	Delay in getting the 'no objection certificate'

Table 3. Feature importance from Select From Model based on random forest model. Features are ordered according to their importance score.

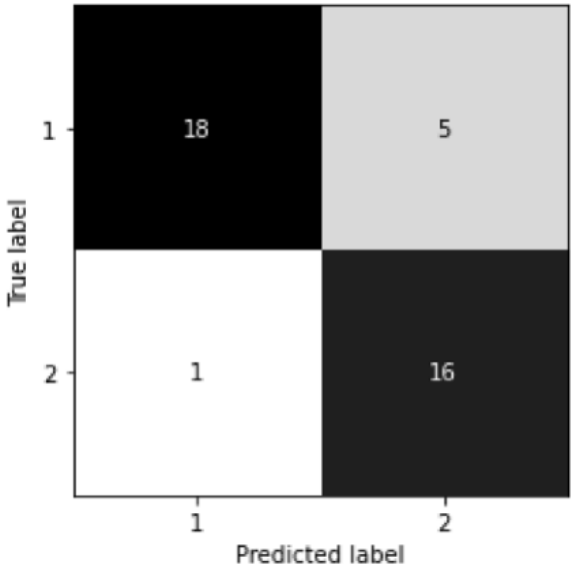

Figure 4. Confusion matrix results based on the random forest model with the SFM feature optimiser (1 for the
rare class and 2 for the often class).

characteristics were chosen from the SFM output. SFM determined that, out of 44 characteristics, are useful for making accurate predictions.

Furthermore, the confusion matrix generated by applying the random forest model with the SFM feature optimiser to the test data is shown in Fig. 4. A total of positive results, 5 negative results, 1 false positive, and negative results were found. Test phase accuracy is therefore calculated as (+) /(+ 5 + 1 +) = 85.00%.

Random forest using the SFM optimizer yields the top 10 features or variables shown in Figure 5. In order to determine the significance of these factors, we calculated the average reduction in impurity as a measure of feature relevance. The significance of each feature is calculated in Mean Decrease in Impurity as the sum over the number of splits that contain the characteristic in relation to the number of samples it splits71. The chart shows that the delays in decision marking attribute contributed the most to the random forest algorithm's classification performance, followed by the cash flow issue attribute and the construction cost underestimation attribute. These top 10 elements were also identified as major drivers to project cost in recent studies of building projects.

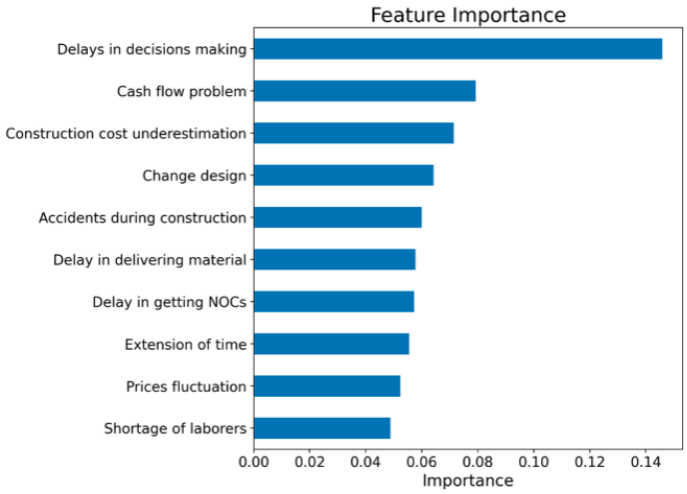

Figure 5. Feature importance (top-10 out of) based on the random forest model with the SFM feature optimiser.

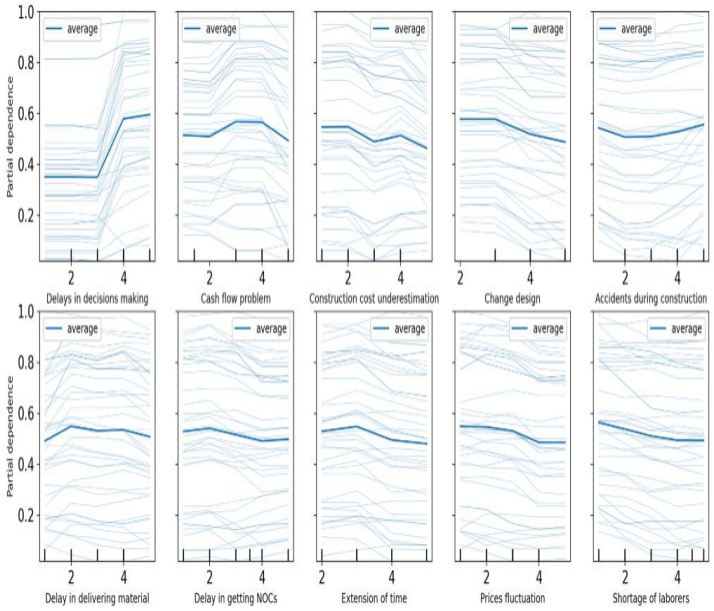

Figure 6. The result of the sensitivity analysis from the partial dependency plot tool for the ten most important features.

Conclusion from case study :

We provided an example of how the suggested research methodology based on machine learning may be used to categorise building projects. With the test dataset, RF's prediction accuracy was the greatest. With a probability of 85.00%, RF can properly identify the class (rare or frequently) of a new data instance that contains information for its characteristics but has not received any clas-ification information. Machine learning algorithms may be trained to be more accurate and efficient at project categorization if given additional data beyond the 139 occurrences of the case study. If we add 100 new data examples, for instance, these algorithms will have an extra 50 instances with which to train, resulting in a 70:30 split.

Because of this capability for ongoing improvement, machine learning algorithms have gained an advantage over more conventional approaches. Some recent research investigate what causes projects to go over budget or over time. Research data analysis was often performed using factor analysis or a closely comparable statistical technique. As a result of applying the suggested machine learning-based framework to the case study, we were able to determine not only which qualities were most essential but also how they ranked and how removing less important components affected the prediction accuracy.

Through the use of GitHub, a website for storing software, we made available the Python code used to build the four machine learning algorithms discussed in this case study. Access the user-friendly version of this programme at https://share.streamlit.io/haohuilu/pa/main/app.py. The hyperparameter settings of the related machine learning algorithms may cause the accuracy results from this link to vary somewhat from one run to the next.

Conclusion:

Machine learning's potential uses in project analytics remain a work in progress. Its employment in the construction industry so far has been restricted to a narrow set of situations, most of which centre on either increasing profits or improving the aesthetics of buildings. Here, our study contributed significantly by developing a machine learning-based approach to deal with issues in project analytics research. We also provided an illustration of how this framework might be used in the context of managing building projects.

This study, like all others, has a few caveats that open up potential avenues for more investigation.

In the first place, there are several cutting-edge machine learning methods (such as handling data-asymmetry problems and estimating kernel density) that are missing from the framework. Second, in order to demonstrate the usefulness of the suggested framework, we only analysed a single case study. Proof of this framework's versatility would come from case studies depicting its use in a variety of project settings. Finally, not all machine learning models and performance measurements mentioned in the literature were taken into account in our investigation. Case in point: while using the recommended research approach, we ignored the Nave Bayes model and accuracy measure.

References:

1. Frame, J. D. & Chen, Y. Why Data Analytics in Project Management? (Auerbach Publications, 2018).
2. Kanakaris, N., Karacapilidis, N., Kournetas, G. & Lazanas, A. In: International Conference on Operations Research and Enterprise Systems. 135–155 Springer.
3. Jordan, M. I. & Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. Science 349, 255–260 (2015).
4. Shalev-Shwartz, S. & Ben-David, S. Understanding Machine Learning: From Theory to Algorithms (Cambridge University Press, 2014).
5. Rahimian, F. P., Seyedzadeh, S., Oliver, S., Rodriguez, S. & Dawood, N. On-demand monitoring of construction projects through
   a game-like hybrid application of BIM and machine learning. Autom. Constr. 110, 103012 (2020).
6. Sanni-Anibire, M. O., Zin, R. M. & Olatunji, S. O. Machine learning model for delay risk assessment in tall building projects. Int. J. Constr. Manag. 22, 1–10 (2020).
7. KB, Sai Sanjana, and Srikanth DV. "Thermal Analysis of Advanced IC Engine Cylinder." *International Journal of Automobile Engineering Research and Development (IJAuERD) ISSN (P)* (2016): 2277-4785.
8. Cong, J. et al. A machine learning-based iterative design approach to automate user satisfaction degree prediction in smart product- service system. Comput. Ind. Eng. 165, 107939 (2022).