

Cost Overrun Prediction Model for Petrochemical Projects in Iran

*Mohammad Mohseni¹, Ernawati Mustafa Kamal¹ and Amin Akhavan Tabassi²

Published: 31 December 2025

To cite this article: Mohammad Mohseni, Ernawati Mustafa Kamal and Amin Akhavan Tabassi (2025). Cost overrun prediction model for petrochemical projects in Iran. *Journal of Construction in Developing Countries*, 30(Supp. 2): 89–117. <https://doi.org/10.21315/jcdc.2025.30.s2.5>

To link to this article: <https://doi.org/10.21315/jcdc.2025.30.s2.5>

Abstract: Cost overrun classification is a critical aspect of project management in Iran's petrochemical industry, demanding accurate predictive models to inform decision-making and mitigate risks effectively. This article focuses on developing a decision tree-based predictive model specifically tailored to classify cost overrun levels as high, moderate and low. Machine learning techniques extract transparent and interpretable rules for cost overrun classification using a rule-based approach within the decision tree framework. The process involves meticulously extracting rules from the decision tree model to provide actionable guidelines based on locally relevant risk factors. The results demonstrate the model's effectiveness in systematically assessing risk factors and accurately predicting cost overrun classifications. Thirteen rules are extracted, each constituting a unique combination of conditions leading to a specific classification outcome, enhancing the model's validity and reliability. The successful completion of this phase delivers a valuable predictive tool for project managers and stakeholders in Iran's petrochemical industry. The developed decision tree model allows stakeholders to navigate cost overrun challenges effectively and optimise project outcomes by enabling informed decision-making and proactive risk management.

Keywords: Prediction model, Cost overrun, Risk management, Petrochemical, Machine learning

INTRODUCTION

The petrochemical industry is a crucial sector in Iran's economy, contributing significantly to its gross domestic product and employing a considerable proportion of the workforce (Ghazizade et al., 2021). Iran's petrochemical projects are primarily concentrated in key regions due to the country's rich natural resources, strategic location and well-established infrastructure. Khuzestan, known for its abundant oil and gas reserves, is a vital hub for the petrochemical sector, hosting numerous projects that leverage its energy wealth. Key areas such as the Petrochemical Special Economic Zone at Mahshahr are critical, attracting many projects due to their proximity to the South Pars Gas Field, one of the world's largest natural gas fields in the Persian Gulf (Shahidi-Nia, 2022). However, petrochemical projects face cost overruns due to several factors, such as management, financial fluctuations in commodity prices, geopolitical tensions, regulatory hurdles and technical challenges (Johari, Cheraghi and Ardakani, 2018). Predicting cost overruns

¹School of Housing, Building and Planning, Universiti Sains Malaysia, 11800 USM, Pulau Pinang, MALAYSIA

²Business School, Manchester Campus, Manchester Metropolitan University, UNITED KINGDOM

*Corresponding author: Mohseni.uni@gmail.com

in petrochemical projects is crucial for effective implementation and advancement. Accurate predictions enable better budgeting, resource allocation and risk management, ensuring projects stay on track and within budget. This proactive approach minimises cost overruns and delays, contributing to the overall success and sustainability of the petrochemical industry (Ghazal and Hammad, 2022). Developing a reliable prediction model for cost overruns in petrochemical projects requires a comprehensive understanding of the factors influencing project costs and their interdependencies. Data-driven approaches can be instrumental in constructing such models by incorporating historical project data into the Iranian context (Liravinia, Modiri and Hafshjani, 2023). Machine learning techniques such as decision tree algorithms offer powerful data analysis tools. The primary research objective is to identify risk factors of cost overruns in petrochemical projects in Iran based on interviews and questionnaires and the second objective's main aim is to develop a reliable prediction model for cost overruns through machine learning analysis. Construction projects in Iran often face cost overruns and petrochemical projects are no exception (Derakhshanalavijeh and Teixeira, 2017).

Petrochemical projects are complex and large-scale and require significant financial resources (Ghazizade et al., 2021). Despite substantial investments in oil, gas and petrochemical projects each year, a staggering 78% of these initiatives exceed their initial budget estimates by more than 25%. Such overruns disrupt timelines and strain financial resources, severely impacting project performance. The persistent cost overruns in Iran's petrochemical industry have profound implications for the government and stakeholders (Esmaeili and Kashani, 2022). Addressing the root causes of cost overruns in construction projects in Iran requires an in-depth comprehension of the various explanations and factors that lead to cost overruns. Conducting a thorough investigation of various construction projects is imperative to achieving a comprehensive understanding and making meaningful comparisons of the causes of cost overruns across different regions of Iran (Amini et al., 2023). Additionally, assessing and prioritising the significant factors contributing to cost overruns within Iranian construction projects can improve construction processes and successfully mitigate such cost overruns. While researchers identified several factors and models for predicting cost overruns, a notable gap exists in understanding the unique challenges specific to the petrochemical industry in Iran (Ashtari et al., 2022).

Investigating cost overruns in the petrochemical sector is crucial to maintaining a sustainable Iranian economy. Identifying and addressing the factors contributing to cost overruns in this sector is essential. According to statistical analysis, estimates of cost overruns in different construction projects are highly inaccurate and biased, highlighting the urgent need to address these biases to make effective decisions (Flyvbjerg and Bester, 2022). The inaccuracies in cost estimations are a significant concern and

the construction industry has acknowledged for some time that traditional cost estimation methods are fraught with uncertainty. Therefore, significant improvements in prediction accuracy are needed (Ghazal and Hammad, 2022). Considering these challenges, developing user-friendly tools for machine learning implementation in construction project management is necessary (Liravinia, Modiri and Hafshjani, 2023). Such tools could enhance cost overrun predictions, improve project performance, reduce financial risks and create more successful construction projects (Sardroud et al., 2021). Based on the issues mentioned earlier, this study aims to develop a prediction model of cost overrun based on the critical risk factors of cost overrun in petrochemical projects in Iran through a decision tree rules-based classifier. Finally, the decision model predicts the percentage of cost overrun in the petrochemical projects in Iran.

LITERATURE REVIEW

Cost overruns in civil projects, particularly in the oil, gas and petrochemical industry, have become increasingly complex, with significant implications for governments, stakeholders and the public. Flyvbjerg, Skamris Holm and Buhl (2003) have found that approximately 86% of projects experience cost overruns, with an average overrun of 28%. The data collected from over 250 large infrastructural projects across different countries and periods demonstrate the consistent occurrence of cost overruns. Furthermore, the study by Gebrehiwet and Luo (2017) supports these findings, stating that construction projects often experience time overruns of 70% and cost overruns of 50%. Bentil et al. (2017) investigated cost overruns in Ghana, revealing a 75% overrun rate for public projects and 34% for private projects. Senouci, Ismail and Eldin (2016) explored cost overruns in the construction sector projects in Qatar, noting a 54% overrun rate.

In Iran, Heravi and Mohammadian (2021) assessed project costs, categorising them into three tiers: (1) small, denoting projects valued at less than USD160,000, (2) medium, encompassing projects valued between USD160,000 and USD480,000, and (3) large, representing projects valued over USD480,000. These thresholds were established considering typical value ranges for urban construction projects in Iran's construction industry. Their analysis revealed that 43.8% experienced cost overruns exceeding 25.0% among large projects, while 56.2% encountered cost overruns of less than 25.0%. According to recent research by Amini et al. (2023), construction projects in Iran are experiencing cost overruns exceeding 25%. Assessment of factors contributing to cost overruns in civil projects is intricate and varied. In a recent study, Ibrahim, Yasin and Hassan (2024) address the specific context of Somalia's construction industry, particularly focusing on the capital city of Mogadishu. Their study examines the primary causes of cost overruns in projects, emphasising the

need for a comprehensive understanding of materials, project management, contractor oversight and unforeseen ground conditions to devise effective strategies for mitigation.

However, Omran, Saleh and Gebril (2023) evaluated the causes of cost overruns in Malaysian construction projects, highlighting issues such as lack of experience and errors in project documentation. Their study provides theoretical support and insights into cost overrun management in the construction industry, pinpointing top factors such as lack of project management experience, delays in information issuance, contractual claims, improvements to standard drawings and errors in bills of quantities and quantity take-off. In addition, Tayyab et al. (2023) discovered a total of 70 factors contributing to cost overruns in high-rise construction projects in India. They identified the 10 most significant reasons, including frequent change orders, construction delays, material price escalation, market inflation, rework, design modifications, erroneous project assessments, unanticipated ground conditions, inaccurate quantity take-off and delayed payments by owners. Moreover, Sohu, Ansari and Jhatial (2021) identifies critical risk factors of cost in the construction sector of Pakistan and ranks causes of cost overruns according to 33 common factors, including 11 financial challenges and troubles faced by contractors. Iranian projects are not exempt from cost overrun.

Among the studies conducted in Iran, Derakhshanlavijeh and Teixeira (2017) evaluated the relative importance of the causes of cost overrun in the gas and oil industry. Their findings revealed that the main causes of cost overrun in this industry include inaccurate cost estimations, improper planning, frequent design changes, inadequate labour/skill availability and inflation of costs of machinery, labour, raw material and transportation prices. Abbasi et al. (2020) identified eight main groups of risk factors for cost overrun in construction projects in Iran, including contractor, owner, design, procurement, equipment, consultant, labour and miscellaneous. Key causes of cost overrun include financial difficulties, poor cash flow, delayed payments, inflation, material price increases, unrealistic schedules and bidding system flaws. Table 1 shows the root risk sources of cost overrun in construction projects, categorised into six groups, along with references from various authors.

Table 1. Previous studies identifying risk sources leading to cost overruns

Source	Management	Coordination and Communication	Contractor and Construction Risk	Environmental Risks	Financial Risks	Stakeholder and Political Risks
Abusafiya and Suliman (2017)				✓		
Ahady, Gupta and Malik (2017)	✓	✓			✓	✓
Akhund et al. (2019)			✓			✓
Allahaim and Liu (2015)	✓			✓		✓
Amini, Rezvani and Tabasi (2021)	✓				✓	✓
Balaji and Venugopal (2020)	✓	✓	✓			
Balali, Zalavadia and Heydarian (2020)	✓			✓		✓
Dabirian, Khanzadi and Abdollahi (2020)	✓		✓	✓	✓	
Derakhshanalavijeh and Teixeira (2017)	✓		✓	✓		✓
Durdyev, Ismail and Abu Bakar (2012)			✓		✓	
Gupta and Kumar (2020)	✓					
Jadhav, Konnur and Patil (2020)				✓	✓	
Johnson and Babu (2018)			✓		✓	
Kamaruddeen, Chang and Wahi (2020)	✓				✓	✓
Keng, Mansor and Ching (2018)		✓	✓	✓	✓	✓
Mahamid (2014)	✓					

Recent studies by Farzanegan and Batmanghelidj (2023) highlight the challenges Iran’s oil, gas and petrochemical sectors face due to economic sanctions. These sanctions, particularly those imposed after 2006, have significantly reduced Iran’s oil export revenues, leading to currency depreciation, high inflation rates and declining output growth (Laudati and Pesaran, 2023). Accurate cost estimation and prediction of cost overrun are essential aspects of project management in construction (Ashtari et al., 2022). Aung et al. (2023) delve into the potential of machine learning algorithms for cost overrun prediction, highlighting their superiority over traditional methods. Similarly, Ghazal and Hammad (2022) contribute to advancing cost overrun prediction models by utilising data mining techniques. Their emphasis on simplicity and effectiveness through clustering, feature selection and classification underscores the importance of practical and efficient approaches in addressing cost overrun challenges.

Eltoukhy and Nassar’s (2021) research is particularly noteworthy for its practical implications in predicting and assessing the severity of cost overruns. Their use of linear regression and machine learning models not only enables accurate prediction of overruns but also provides a means to assess their severity. This research offers a valuable tool for stakeholders in the construction industry, allowing them to better anticipate and manage cost overruns. Table 2 summarises and comprehensively lists ways to predict cost overruns and develop cost estimation models. It provides the method and algorithm of a prediction model with a specific approach. This table is compiled from previous studies between 2019 and 2023.

Table 2. Summarising studies of cost overrun prediction model

Authors	Year	Method	Approach
Aung et al.	2023	Random forest	Cost predicts of public construction projects
Leu, Liu and Wu	2023	Machine learning	Prediction potential of machine learning
Moghayeddi and Windapo	2022	The adaptive neuro-fuzzy	Modelling the uncertainty of cost and time in highway projects
Nassar and Erzaij	2022	Multiple linear regression	Developing models to predicting the effect of crises on construction projects
Shoar, Chileshe and Edwards	2022	Random forest regression model	Cost overruns prediction in high-rise residential building projects
Sharma et al.	2022	Machine learning	Cost prediction and optimisation in construction
Mrad, Titirla and Larbi	2021	Linear regression and neural network models	Comparison of linear regression and neural network models to estimate cost

(Continued on next page)

Table 2 *Continued*

Authors	Year	Method	Approach
Srinivasan et al.	2021	Microwave background, artificial neural network (ANN) and multiple regression analysis	Cost prediction model in construction projects
Alemu	2022	Regression model	Time and cost prediction model for public building projects
Monteiro et al.	2021	Multiple linear regression	Cost modelling from the contractor perspective
Balta, Dikmen and Birgonul	2021	Bayesian belief network	Predicting and mitigating delay and cost
Mahdavian et al.	2021	Machine learning algorithms	Predictive modelling of highway construction cost
Plebankiewicz and Wieczorek	2020	Fuzzy Mamdani	Cost overrun risk prediction model
Gotlur et al.	2020	Neural networks	Review on prediction of cost overrun in construction
Ragheb, Massoud and Hamisa	2020	ANN model	Developing an ANN model to predict the actual cost overruns
Ghazal and Hammad	2022	Knowledge discovery in databases techniques	Cost overruns prediction model
Shiha, Dorra and Nassar	2020	Neural networks	Model for prediction of construction material prices
El-Dash, Ramadan and Youssef	2019	Regression model	Prediction models for construction projects in the Middle East
Tijanć, Car-Pušić and Šperac	2019	ANN	Cost estimation in road construction
Petrusheva, Car-Pušić and Zileska-Pancovska	2019	Support vector machine	Model for prediction of road structures construction costs

The conceptual framework in this research outlines a structured approach to predicting cost overrun classifications using risk factors as input variables, as shown in Figure 1. The conceptual framework illustrates that the process layer develops predictive models for analysing cost overruns in petrochemical projects through data preprocessing, factor selection, model training with historical data and cost overrun prediction. The output layer translates these predictions into actionable insights by categorising them into high, moderate and low-cost overruns and evaluates the model's performance to ensure reliability and effectiveness.

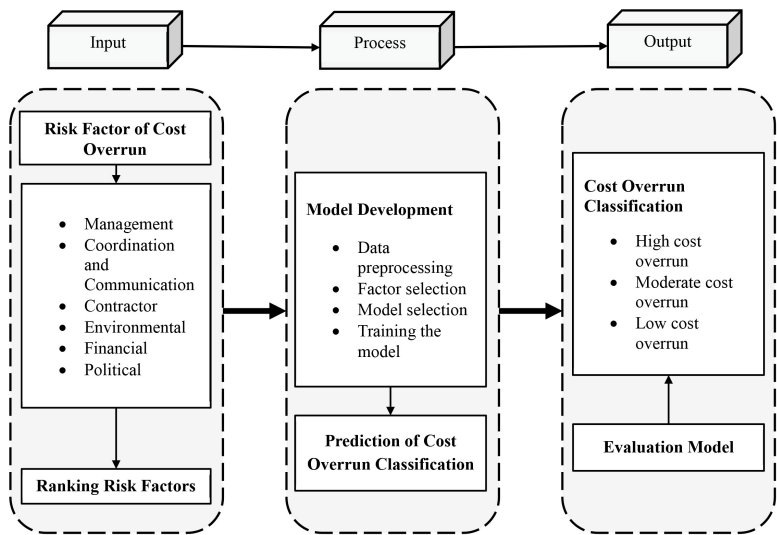


Figure 1. Conceptual framework

RESEARCH METHODOLOGY

The mixed-methods approach was chosen for this research due to its alignment with the aims and objectives, offering several advantages. The research objectives include: (1) to identify risk factors for cost overruns and (2) to develop a reliable prediction model for cost overruns. By expanding on the findings of each phase in subsequent phases, the study facilitates a continuous exchange between qualitative insights and quantitative validation (Creswell and Clark, 2017). Specifically, this approach allows for a comprehensive analysis by combining quantitative data, which provides statistical rigor and generalisability, with qualitative insights that offer depth and context. By integrating these methods, the research can address the complexity of the issues at hand more effectively, ensuring a robust understanding of the factors contributing to cost overruns in petrochemical projects. Moreover, mixed methods facilitate the triangulation of data, enhancing the validity and reliability of the findings. This approach is particularly suitable for exploring multifaceted problems in construction project management, where both numerical data and nuanced stakeholder perspectives are crucial.

Data Collection

The initial step in identifying risk factors for cost overruns in petrochemical projects in Iran involved conducting exploratory interviews with five professionals with diverse experiences and job roles, both face-to-face

and online via Skype. Qualitative methodologies accept small samples, prioritising depth over breadth, aligning with the selection of five participants. This approach mitigates the risk of superficial data often associated with larger samples, guaranteeing comprehensive and meaningful insights for the research objectives (Rajput and Agarwal, 2020). This study used non-probability purposive sampling to select participants based on their knowledge and experience specifically related to the risk factor of cost overrun in petrochemical projects. These discussions, guided by 62 risk factors identified from the literature, refined 30 key risk factors for the subsequent phase.

Purposive sampling techniques are commonly utilised in research papers across various research paradigms. This technique helps to ensure the selection of a quality sample without biases, thereby increasing the reliability of the study's findings (Nyimbili and Nyimbili, 2024). Following the interviews, a questionnaire was conducted to collect quantitative data on the identified risk factors. The survey was distributed among 112 industry experts in Iran and yielded 99 responses. The experts were asked to provide their professional opinions and evaluations on the frequency and severity of cost overruns. The research concentrated on the Mahshahr Special Economic Zone, a pivotal industrial area for petrochemical projects in Iran, ensuring contextual relevance. Employing purposive sampling, participants were selected based on their expertise in petrochemical projects. These individuals are involved in organisations as contractors, consultants, financial experts, owners and project managers. The questionnaire consisted of three sections, gathering demographic details and project experiences while assessing risk factors' frequency and severity using a 5-point Likert scale. This helped standardise analysis and allowed for benchmarking against prior studies for comparative insights and trend analysis (Kamaruddeen, Chang and Wahi, 2020; Tayyab et al., 2023; Ibrahim, Yasin and Hassan, 2024). In the last step, statistical tools such as Microsoft Excel (Microsoft Corporation, Washington, USA), Statistical Package for the Social Sciences (SPSS) version 26 (IBM Corporation, New York, USA) and Python IDLE version 3.11 (Python Software Foundation, Delaware, USA) were used to verify and analyse the collected data.

Data Reliability

In this research, evaluating the questionnaire's reliability necessitates the calculation of Cronbach's alpha, a measure critical for evaluating the consistency and reliability of responses collected using a Likert scale (Johnson and Babu, 2018). This approach assesses reliability and contributes to the validation of the response scores. Cronbach's alpha is a statistical measure calculated from a dataset using the SPSS software. This measure determines how closely related a set of items is when considered a group. It assesses the degree to which the dataset's items measure the same underlying construct or concept. In other words, it indicates the reliability

of the dataset by measuring the extent to which all the items in the dataset are measuring the same thing. The test results in Table 3 indicate a high level of internal consistency among the items in each dataset. Generally, a Cronbach’s alpha value above 0.7 is considered acceptable, values above 0.8 are considered good and values above 0.9 are considered excellent. Both datasets demonstrate good internal consistency. These findings validate the dataset’s suitability for further in-depth analysis.

Table 3. Cronbach’s alpha reliability test results

Measure	Value
Severity data	0.884
Frequency data	0.857
Cronbach’s alpha (total)	0.870

Dataset and Preprocessing

This study used the risk importance index (RII; Equation 1) to assess and prioritise risk factors in petrochemical projects, aligning with the Project Management Institute’s (2017) recommendations for quantitative risk analysis methods. The RII method, essential for ranking risk factors, evaluates each factor’s frequency index (FI; Equation 2) and severity index (SI; Equation 3). Ammar, Abdel-Monem and EL-Dash (2022) highlighted the necessity of quantitative techniques like RII in construction projects, emphasising its broad verification and common use in risk management, particularly in construction and engineering. Table 4 offers an in-depth understanding of RII scores for 30 risk factors, highlighting those with a more significant effect on causing cost overruns in petrochemical projects.

$$RII = FI \times SI$$

Eq. 1

$$FI = \frac{1}{N} \sum_{i=1}^N r_i$$

Eq. 2

$$SI = \frac{1}{N} = \sum_{i=1}^N r_i$$

Eq. 3

Where, represents the total number of observations or responses. It is the denominator of the fraction, which normalises the sum of ratings to compute an average. Next, is the summation symbol, indicating the sum of the ratings or scores (r_i) for each observation from to N , where i indexes each observation; r_i is the rating or score of the i th observation. These ratings are the individual assessments or measurements for each event considered. In this research, “Very Low” = 1, “Low” = 2, “Moderate” = 3, “High” = 4 and “Very High” = 5.

Table 4. Ranking cost overrun risk factors in the petrochemical projects

ID	Risk Factor	SI	FI	RII	Rank
FIN-005	Inflation	4.4	3.8	16.6	1
MGT-005	Poor project management	4.3	3.6	15.6	2
POL-002	Economic sanctions	4.3	3.6	15.4	3
CON-002	Complexity of design and construction	4.2	3.5	15.0	4
FIN-004	Currency exchange rates	4.2	3.5	14.8	5
MGT-001	Lack of experience in project management	4.2	3.5	14.7	6
CON-001	Lack of experience of contractors	4.2	3.5	14.5	7
POL-003	Corruption and bribery	4.2	3.4	14.3	8
CON-006	Project size	4.1	3.5	14.2	9
CON-007	Design errors and additional work	4.1	3.5	14.1	10
MGT-003	Inadequate risk assessment	4.1	3.4	13.9	11
POL-001	Changes in government policies and regulations	4.0	3.4	13.5	12
COM-001	Conflict between project parties	4.0	3.3	12.9	13
MGT-004	Poor contract management	3.9	3.2	12.5	14
POL-005	Delayed decision-making and approvals	3.8	3.1	12.0	15
CON-008	Delay in material delivery	3.8	3.1	11.9	16
CON-005	Supply chain inefficiencies	3.8	3.1	11.7	17
FIN-001	Complexity of managing multiple sources of funds	3.8	3.0	11.4	18
CON-009	Unforeseen difficulties in transportation	3.7	2.9	10.6	19
MGT-002	Inadequate project monitoring and planning	3.6	2.8	10.0	20
POL-004	Inconsistent application of laws and regulations	3.5	2.7	9.6	21
ENV-001	Natural disasters	3.4	2.7	9.0	22
CON-004	Shortage of skilled labour	3.3	2.5	8.4	23
ENV-003	Air pollution and climate change	3.1	2.6	8.0	24
FIN-003	Owner fund shortage and payment delays	3.2	2.5	7.9	25
CON-003	Poor quality control	2.9	2.3	6.6	26
FIN-002	Contractor fund shortage	2.8	2.1	5.8	27
COM-003	Lack of coordination	2.6	2.0	5.3	28
COM-002	Lack of communication	2.5	2.0	5.0	29
ENV-002	Soil and groundwater conditions	2.5	1.9	4.8	30

Characterising Petrochemical Cost Overrun

A comprehensive survey, which involved gathering information through a questionnaire, covered a wide range of project details and experiences with cost overruns in the petrochemical industry. This in-depth section aimed to provide a holistic view of the percentage of cost overruns in petrochemical projects. The survey’s categorisation of private and public projects into three groups based on their cost overrun, namely, low (below 15% overrun), moderate (15% to 30% overrun) and high (above 30% overrun), was based on a thorough analysis of the data. Out of the 25 projects surveyed, 6 in the public sector and 3 in the private sector fell into the low-cost overrun category. The moderate cost overrun category included two public-sector projects and four private-sector projects. Lastly, six public-sector and four private-sector projects are categorised as high-cost overrun projects, providing a comprehensive picture of the industry’s cost overrun landscape. Table 5 outlines three classes of cost overrun outcomes. This categorisation allowed for the development of a robust predictive model to classify cost overruns in petrochemical projects accurately.

Table 5. Ranking cost overrun risk factors in the petrochemical projects

Category	Description	Number of Projects
Low	Cost overrun less than 15%	9
Moderate	Cost overrun between 15% and 30%	6
High	Cost overrun more than 30%	10
Total		25

Machine Learning Method

This research implemented unsupervised and supervised machine learning techniques. The unsupervised approach efficiently identified the most critical risk factors as independent variables for the prediction model dataset. Furthermore, the supervised learning methodology developed a cost overrun prediction model. In addition, all the analyses were conducted using Python IDLE and Jupyter Notebook (<https://jupyter.org/>). This study utilised K-means clustering, a robust unsupervised machine learning technique, to identify natural thresholds between risk factor categories. Inspired by the recommendations of Mittal, Sharma and Singh (2019) and Ikotun et al. (2023), K-means clustering is chosen for its efficacy in clustering data, particularly in critical datasets such as risk factors. The centroids derived from the clustering process serve as meaningful indicators for delineating these levels, providing a data-driven and statistically robust approach to risk factor categorisation. This research adopted the K-means formula as shown in Equation 4.

$$J = \sum_{j=1}^k \sum_{x_i \in S_j} \|x_i - c_j\| \quad \text{Eq. 4}$$

Where J is the within-cluster sum of squares and I_{ij} is an indicator functions equal to 1 if observation i is in cluster j and 0 otherwise and x_i equals the i th observation. S_j is the set of all observations assigned to the j . Next, c_j is the centroid of cluster j and $\|x_i - c_j\|$ is the Euclidean distance between x_i and c_j . Determining the optimal number of clusters is an essential step in the K-means clustering algorithm. In this analysis, the number of clusters, denoted as k , was set to 5 to determine risk thresholds. This decision was informed by categorising the risk factors into five categories (“Very Low”, “Low”, “Moderate”, “High” and “Very High”). This method divides risk factors into distinct clusters according to their RII values. Each cluster is represented by a centroid that serves as a reference point for its corresponding risk level.

Decision Tree

The decision tree is one of the powerful algorithms for supervised machine learning techniques. Decision tree algorithms, such as ID3, C4.5 and CART, are commonly used for data classification (Bishnoi and Hooda, 2022). Based on the aim of this research, this model was utilised to predict the level of cost overrun in the petrochemical projects in Iran. The algorithm is suitable for this research, as Xu (2023) suggested, as it enables numerical and categorical data inputs. CART is particularly valued for its clarity and effectiveness, employing binary splits to reduce impurity. Pruning is utilised to avoid overfitting and attribute selection uses information theory, particularly entropy, formulated as Equation 5, to measure data impurity and improve predictive precision (Piryonesi and El-Diraby, 2020).

$$H(S) = -\sum_{i=1}^k p_i \log_2(p_i) \quad \text{Eq. 5}$$

Where, $H(S)$ represents the entropy of a set S , n is the number of classes and P_i is the proportion of the elements in S that belong to class i . This measure quantifies the uncertainty or disorder present in a dataset, with higher values indicating more heterogeneity among the data. Entropy serves as a foundation for the concept of information gain, a criterion for selecting the attribute that best splits the data, defined as:

$$IG(S, A) = H(S) - \sum_{\text{div} \in \text{Value } A} \frac{|S_{\text{div}}|}{|S|} H(S_{\text{div}}) \quad \text{Eq. 6}$$

Where, $IG(S, A)$ is the information gain of an attribute A , value (A) represents the set of all possible values for attribute A , S_v is the subset of S for which attribute A has value v and $|S_v|/|S|$ is the proportion of S that ends up in v . This formula calculates the difference in entropy from before to after the dataset is split on attribute A , thus, quantifying the reduction in uncertainty or disorder

achieved by the split. This study aimed to develop a prediction model that categorises cost overruns into low, moderate and high categories based on risk factors. The approach involved extracting rules from a decision tree by creating a path from the root to the leaf node and combining split attributes with the prior. The study utilised decision trees to create a clear set of IF-THEN rules for classification, moving from theory to practical implementation (Li et al., 2017). This transition simplified the decision tree hierarchy and improved its usability for specific class predictions. Figure 2 illustrates the concept of the rule-based decision tree model used in this research.

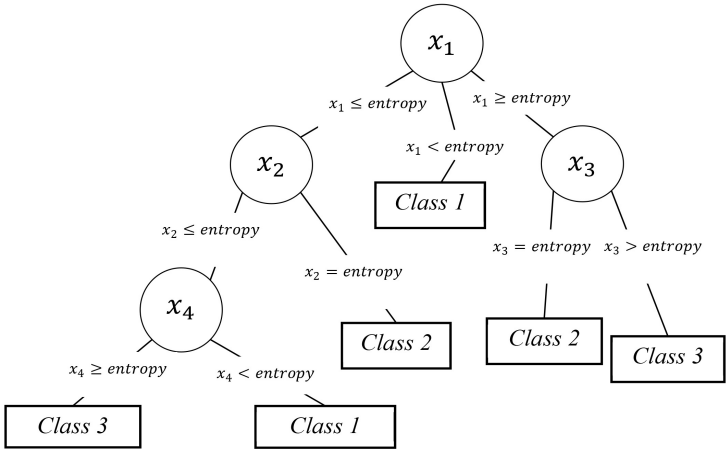


Figure 2. The rule-based decision tree

Evaluation and Validation of Model

This study utilises the confusion matrix and receiver operating characteristic (ROC) curve to assess the effectiveness of a model. The confusion matrix provides a detailed analysis of the model's predictive abilities, evaluating accuracy, precision, recall and F1-score, as shown in Equations 7 to 10. Pencina, D'Agostino and Massaro (2013) suggests employing ROC curves and the AUC metric to evaluate a model's discriminatory power across different threshold levels, illustrating the trade-offs between sensitivity and specificity. By combining these techniques, a comprehensive evaluation of the model's strengths and areas for improvement is achieved. The confusion matrix displays the actual and predicted classifications, deriving performance metrics such as precision, sensitivity and specificity. On the other hand, the ROC curve plots the true positive rate against the false positive rate, with the area under the curve (AUC) providing an aggregate measure of the model's performance.

$$\text{Accuracy} = (TP + TN) / (TP + FN + FP + TN) \quad \text{Eq. 7}$$

$$\text{Precision} = TP / (TP + FP) \quad \text{Eq. 8}$$

$$\text{Recall} = TP / (TP + FN) \quad \text{Eq. 9}$$

$$\text{F1-score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad \text{Eq. 10}$$

Where, TP (true positives) is the number of correct positive predictions, TN (true negatives) is the number of correct negative predictions, FP (false positives) is the number of incorrect positive predictions and FN (false negatives) is the number of incorrect negative predictions.

ANALYSIS AND RESULTS

Analysis of K-means Clustering Risk Factors of Cost Overrun

The K-means clustering is used to analyse the data and determine risk thresholds. The method divided risk factors into clusters based on their RII values, each represented by a centroid. Five predetermined risk categories were present: “Very Low”, “Low”, “Moderate”, “High” and “Very High”. A density plot illustrating the distribution of RII values and the corresponding cluster centroids, which served as boundaries between risk classifications, is shown in Figure 3.

The density plot illustrates the distribution of RII values, highlighting the density and placement of risks. Peaks in the density curves represent clusters of risk factors, with shaded areas indicating their concentration. Five centroids categorise these risks:

1. Centroid 1 (RII = 5.32, Very low).
2. Centroid 2 (RII = 8.78, Low).
3. Centroid 3 (RII = 11.82, Medium).
4. Centroid 4 (RII = 14.29, High).
5. Centroid 5 (RII = 15.85, Very high).

The K-means clustering algorithm grouped risk factors into these five clusters based on similarity, with centroid values detailed in Table 6. The analysis identified 12 significant risk factors critical for predicting cost overruns, with the high and very high categories being particularly important. The high category includes nine risk factors (e.g., MGT-002, POL-003 and CON-009), while the very high category comprises three key factors (POL-002, MGT-005 and FIN-005). Figure 4 visually represents these findings.

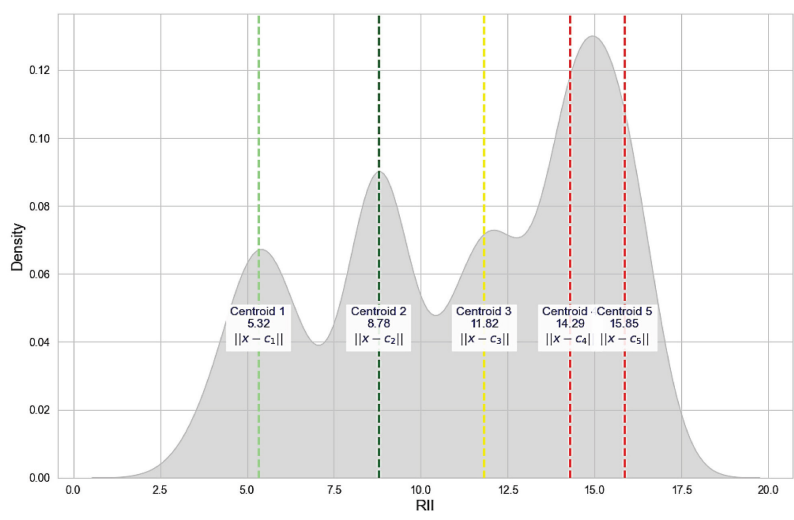


Figure 3. Distribution of RII values and the associated cluster centroids

Table 6. Clustering cost overrun risk factors

K-Means Cluster	Risk Level	Centroid Value	Number of Factors	Risk Factors
1	Very low	5.32	5	ENV-002, COM-002, COM-003, FIN-002 and CON-003
2	Low	8.78	5	FIN-003, ENV-003, CON-004, ENV-001 and POL-004
3	Moderate	11.82	8	MGT-002, CON-009, FIN-001, CON-005, CON-008, POL-005, MGT-004 and COM-001
4	High	14.29	9	POL-001, MGT-003, CON-007, CON-006, POL-003, CON-001, MGT-001 FIN-004 and CON-002
5	Very high	15.85	3	POL-002, MGT-005 and FIN-005

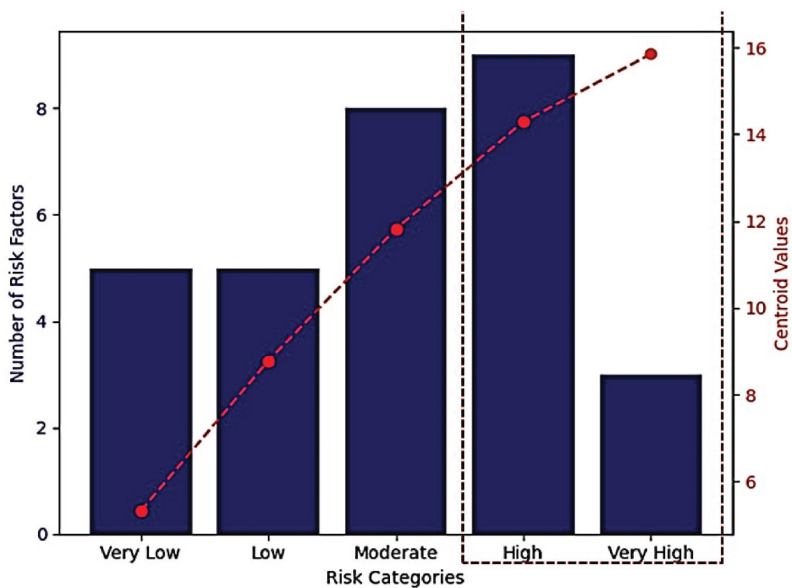


Figure 4. Category of cost overrun levels based on K-means threshold

Data Test and Train for Implementation Model

The decision tree model involves dividing the dataset into training and testing sets. Initially, a baseline decision tree model is created, followed by optimising its performance through hyperparameter tuning. Once the evaluation metrics are considered, the best-performing model is selected and validated against the testing dataset. The final model is then thoroughly evaluated and a feature importance analysis conducted to identify any significant risk factors. Insightful decision rules are extracted from the model and deployed for predictions. The dataset includes a group of independent variables (risk factors of cost overrun) indicated as predictor variables ($x_1, x_2, \dots, x_{12}$); these 12 variables are key risk factors identified through analysis and are essential in predicting cost overrun. Furthermore, the dependent variable (y) was obtained through detailed data collection involving 25 petrochemical projects. The dependent variable, response variable (y), categorises the cost overrun outcomes into three classes (Class 1 = “Low”, Class 2 = “Moderate” and Class 3 = “High”). The process of data splitting involves dividing the dataset into training and testing sets, with a split of 70% for training and 30% for testing. In Figure 5, a bar chart displays the distribution of cost overrun categories in both sets following the split. Each bar represents the count of records for a category (0 for “Low”, 1 for “Moderate” and 2 for “High”), distinguishing between the training and testing datasets. The chart illustrates a consistent distribution of

cost overrun categories in both sets, ensuring that the model can effectively generalise to new data. This balanced split helps maintain the reliability of the model evaluation.

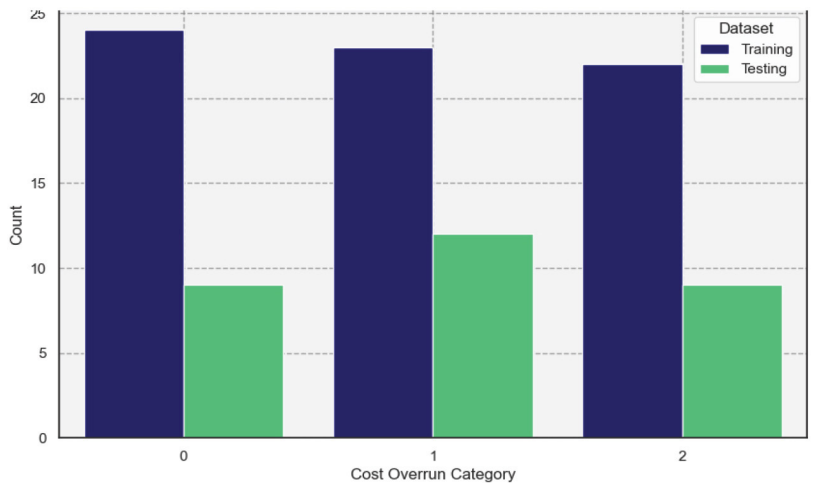


Figure 5. Training and testing datasets

Model Implementation

The initial decision tree algorithm model was trained on a subset of the dataset, with a portion reserved for testing performance evaluation. Metrics such as accuracy, precision, recall and ROC AUC score were used to assess the model’s performance in classifying cost overrun categories. The initial model showed misclassifications in predicting cost overrun, struggling to distinguish between different levels. The ROC AUC score indicated a mediocre discriminative ability but could be improved through hyper-parameter tuning. The ideal settings involved maximum depth without limitations, minimum sample split of 5 and minimum sample leaf of 2 to balance overfitting and underfitting. This optimisation resulted in higher accuracy, precision, recall and F1-score metrics, improving the model’s predictive capability and generalisation. The model’s overall accuracy, as presented in Table 7, is at 0.91, demonstrating excellent performance. With a macro average precision of 0.92, the model exhibits high precision across all classes. Moreover, the macro average recall stands at 0.905, indicating an exceptional ability to identify positive instances. The model’s macro average F1-score of 0.897 further confirms its high classification performance. Impressively, the weighted average precision of 0.93 showcases the model’s consistent high precision, even in the presence of class imbalance. Additionally, the model’s weighted average recall of 0.900 suggests its effective handling of class distribution.

Furthermore, the model achieves a commendable weighted average F1-score of 0.899, illustrating its robust performance regardless of class distribution. Notably, the model excels across all classes, achieving perfect scores for class 2. Its well-calibrated nature, balanced precision and recall result in minimal false positives and false negatives.

Table 7. Overall performance tuned model

Precision	Recall	F1-score	Support
Class 0	0.99	0.714	0.833
Class 1	0.75	1.000	0.857
Class 2	0.99	0.990	1.000
Accuracy	–	–	0.910
Macro	0.92	0.905	0.897
Weighted	0.93	0.900	0.899

Decision Tree Model Validation and Evaluation

Cross-validation analysis demonstrated high accuracy, precision, recall and F1-score levels in 10 folds, confirming the model’s robustness. Metrics declined in folds 6 and 10, providing insights into performance under different data subsets, as shown in Figure 6. High-performance metrics indicate a well-calibrated and reliable model for the intended application. The decision tree classifier performs highly in predicting cost overrun, exhibiting consistency in core metrics across different data segments. Metrics consistently exceed the 0.6 threshold, with the majority surpassing 0.8, indicating reliable performance in complex classification.

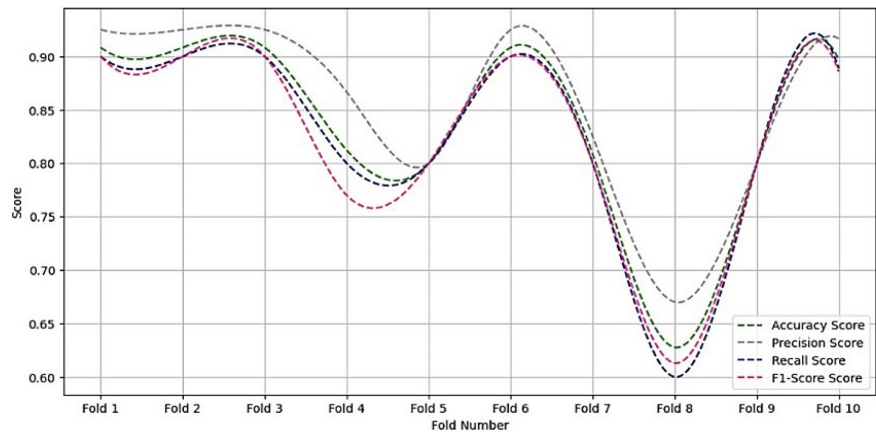


Figure 6. Cross-validation scores

The ROC curve analysis for the decision tree classifier demonstrated excellent performance across all classes, with micro-average and macro-average AUC of 0.92 and 0.93, respectively. Class 2 achieved a perfect AUC of 1.00, indicating exceptional accuracy. Classes 0 and 1 had robust discrimination power, with AUC of 0.86 and 0.93, respectively. The overall shape of the ROC curve confirmed the classifier’s effectiveness. With adjusted decision thresholds, the classifier is well-calibrated and can confidently provide reliable class predictions, as shown in Figure 7.

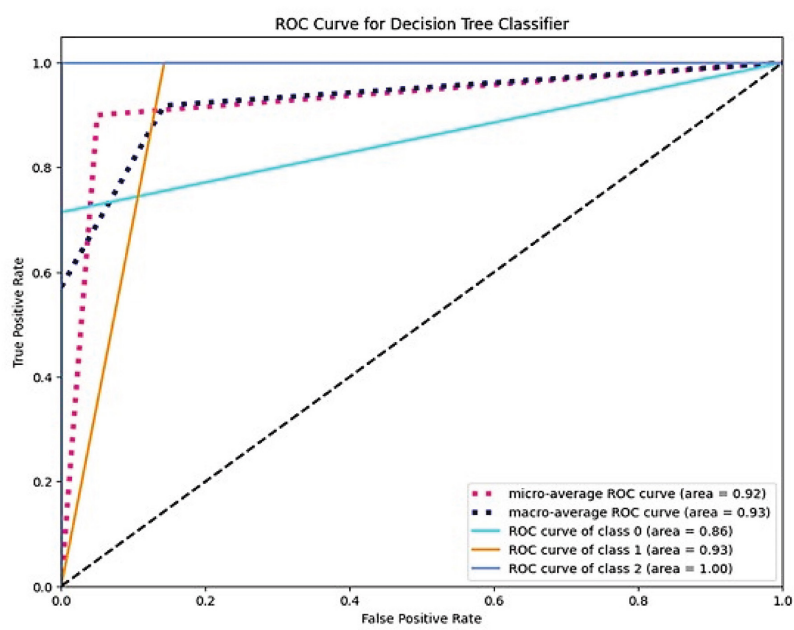


Figure 7. ROC curve for evaluation decision tree model

Decision Tree Model Visualisation

A visualisation of the decision tree classifier prediction model, as depicted in Figure 8, provides an informative illustration of the model’s decision tree process. At the onset of the tree, the primary factor considered by the root node is MGT-005 due to its significant impact on the prediction model; subsequent nodes evaluate additional criteria, including POL-002 and FIN-004. The tree is composed of nodes and branches. Each internal node represents a decision based on one of the factors (e.g., POL-002 and MGT-001) and each branch represents the outcome of that decision, leading to another node. The topmost node (here, testing MGT-005) is the starting point of the decision-making process. According to the model, these factors provide the most significant initial split. At each node, the decision is made based on a

threshold. For instance, at the root, the model checks if MGT-005 is less than or equal to 2.5. This threshold plays a crucial role in determining the path the tree takes, influencing the decision rules inferred from the factors' values. The concept of entropy impurity is key to understanding the tree's predictive capabilities. An entropy impurity of 0 signifies perfect purity, indicating that all elements at that node belong to the same class. This value is derived from the number of observations in the training dataset that reached the node. The distribution of classes for the samples at that node is shown in an array. For example, a value [1, 0, 11] indicates 1 sample of the first class, none of the second and 11 of the third class present at that node. The leaf node at the end of a branch predicts the class based on the path followed down the tree. This decision tree is a model that has been learned from the training data to make predictions by creating a set of decision rules inferred from the factors' values. It is a highly interpretable model, which allows precisely how predictions are made and helps understand the decision logic and the relative importance of features within the model.

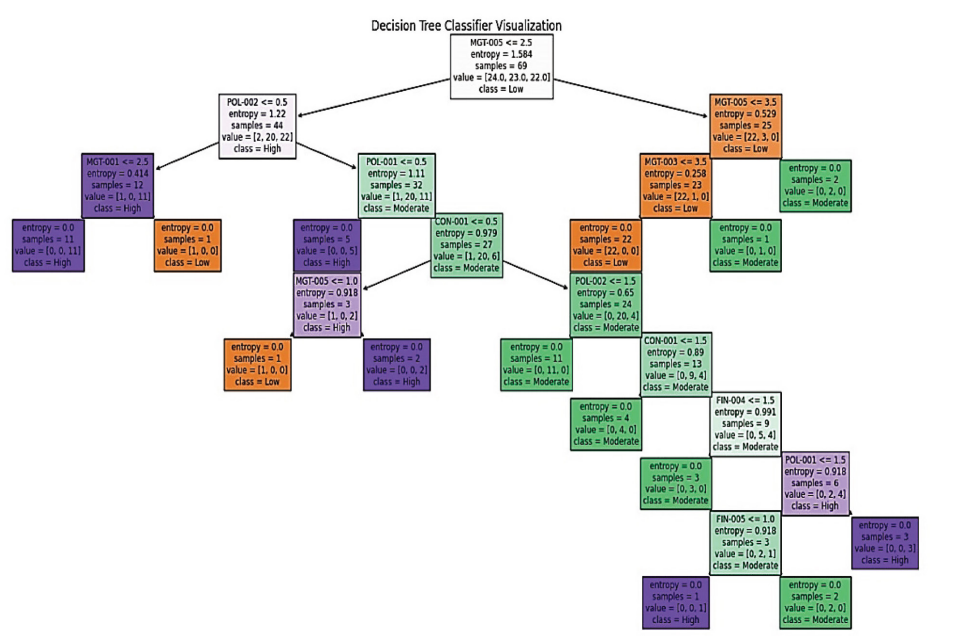


Figure 8. Decision tree visualising the leaves and nodes

Rules Extraction of Decision Tree Prediction Model

The process of extracting rules from the decision tree, aimed at predicting the class of cost overrun as (high, moderate and low), was executed precisely and concluded. The 13 rules extracted provide a transparent and highly

interpretable set of guidelines, each constituting a unique combination of conditions that leads to a specific classification outcome. This meticulous extraction of rules underscores the model’s validity, attesting to its ability to assess risk factors and provide dependable predictions. The rule-based decision tree will provide the final model presented visually, as shown in Figure 9, which is the rules of prediction cost overrun classification. The model takes a set of risk factors as input, processes them through a series of decision rules (rule 1 to rule 13) and outputs a classification into one of three categories: low (cost overrun < 15%), moderate (15% < cost overrun < 30%), or high (cost overrun > 30%). In this model, specific risk factors such as political risks (POL-001, POL-002 and POL-003), construction risks (CON-001, CON-002, CON-006 and CON-007), management risks (MGT-001, MGT-003 and MGT-005) and financial risks (FIN-004 and FIN-005) are inputs. The decision tree evaluates each risk factor, applying a series of IF-THEN logical rules based on the historical data patterns learned during the training phase. The rules effectively dissect the complexities of risk interplay, with each rule likely incorporating conditions or thresholds of specific risk factors to guide the prediction. For instance, one rule might specify that if POL-002 is high and MGT-005 is low, the prediction might be a high cost overrun risk. Finally, the model’s output layer predicts the cost overrun category.

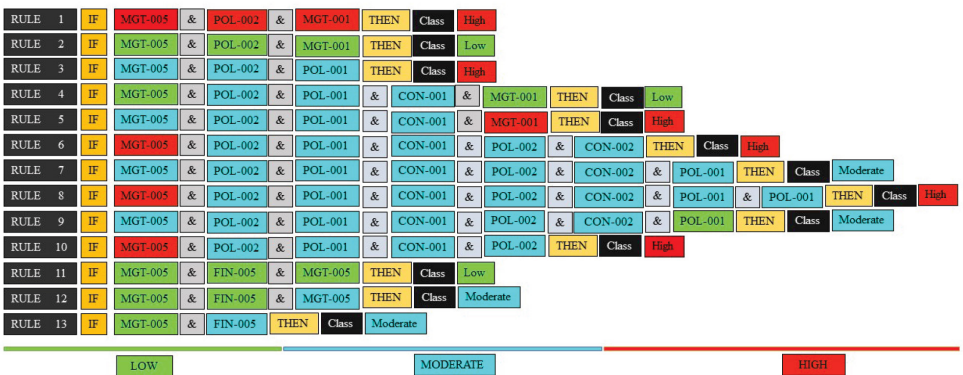


Figure 9. Rules of cost overrun prediction model

DISCUSSION

A comprehensive literature review was conducted to identify the risk factors contributing to cost overruns in petrochemical projects in Iran. This review resulted in an initial list of 62 potential causes, refined into a condensed list of 30 specific risk factors. This refinement process was conducted through collaborative interviews with industry experts, providing valuable insights

into the challenges encountered in Iranian petrochemical projects. The selection of interviewees was strategic, focusing on those with a rich mix of experience, various job roles and extensive project work. This process aimed to list these factors and understand their origins, interactions and impacts on project costs.

Furthermore, an analysis using the RII was conducted to rank the significance of these risk factors through a survey. The survey was distributed among 112 industry experts in Iran and yielded 99 responses. They were asked to provide their professional opinions and evaluations on the frequency and severity of cost overruns. The RII considered the severity and frequency of the 30 identified risk factors. This assessment aimed to identify the risks with the most significant potential to lead to cost overruns, highlighting factors such as inflation, inadequate project management and economic sanctions as the most influential contributors to cost overruns in petrochemical projects in Iran. For the prediction of cost overrun, the initial step involves preparing and cleaning data, including risk factors identified in previous steps and data from 25 petrochemical projects collected through surveys. These projects are categorised into three classes based on the percentage of cost overrun: below 15%, between 15% and 30% and above 30%. Subsequently, a decision tree classification model is developed to predict the level of cost overrun in petrochemical projects. The significance of this study is underscored by the fact that the initial model achieves an accuracy of 75%, which could be better. However, efforts are made to enhance its performance by tuning hyperparameters. Following parameter tuning, the model's accuracy significantly improves, reaching nearly 91%. In previous research, Shoar, Chileshe and Edwards (2022) and Tajziyehchi et al. (2020) revealed that the average accuracy attained with random forest models was approximately 88%, which was deemed acceptable. However, the current study has achieved an accuracy of 91%, surpassing these previous results. Moreover, research by Ghazal and Hammad (2022) found that the random forest model emerged as the best machine learning model, yielding an accuracy of 65.22% and the decision tree model developed by Ashtari et al. (2022) for cost overrun prediction achieved 65.00% accuracy. In contrast, the decision tree model in this study demonstrated significantly higher accuracy, exceeding this result by at least 25%. These comparisons underscore the significance of continuous improvement and advancement in predictive modelling. It is essential to acknowledge that the accuracy of models is greatly influenced by the quality of input data and the refinement of the model development process. Lower and higher accuracy results are valuable, as they drive the iterative process of model enhancement, assuring continual improvement.

CONCLUSION

Cost overruns are a significant issue in the construction industry, particularly in petrochemical projects. Understanding the critical reasons behind cost overruns and their impact is challenging. This study aimed to develop a cost overrun predictive model for petrochemical projects within Iran. The study focused on identifying and ranking risk factors of cost overrun and developing a cost overrun prediction model in petrochemical projects in Iran. A methodological approach was adopted to achieve this, involving interviews, questionnaire analysis and machine learning techniques. As a result, 30 risk factors were identified, significantly impacting cost overruns in petrochemical projects in Iran. Furthermore, machine learning with a decision tree model provides a structured framework for analysing and predicting cost overrun. The model exhibits excellent performance with an overall accuracy of 0.91, high precision (0.92), recall (0.905) and F1-score (0.897). It also maintains high precision (0.93) and recall (0.90) with class imbalance. The model achieves a commendable weighted average F1-score of 0.899 and excels in all classes, particularly class 2, with minimal false positives and false negatives. This decision tree model has been designed to make predictions by creating a set of decision rules based on the values of different factors. These rules effectively break down the complexities of risk interplay, with each rule likely considering specific risk factors' conditions or thresholds to guide the prediction. The research also provides valuable insights for the petrochemical industry in Iran, informing policy, education and training programmes. Additionally, the practical implications of this research can be implemented in real-world practices to improve industry outcomes.

Contribution of Research

The research significantly enhances the understanding of Iran's petrochemical industry by providing valuable insights that benefit policy making, education and project manager training programmes. Meanwhile, the practical implications of applying these insights to real-world research studies are a valuable addition to the existing knowledge base of the petrochemical industry in Iran. It proposes insights that can be useful in developing policies, education and training programmes. In addition, the research contributes to the practical basis for deploying a predictive model developed in the research through a web-based platform or app. By doing so, the research aims to make predictive analytics accessible to project managers and stakeholders in a user-friendly and convenient manner, ensuring a smooth transition to this advanced technology. This practical proposal translates theoretical advancements in predictive modelling into actionable strategies for implementation, ultimately aiming to improve decision-making and project outcomes by providing easy access to predictive insights.

ACKNOWLEDGEMENTS

The authors would like to thank Universiti Sains Malaysia and the School of Housing, Building and Planning for their support. The contributions and collaboration of project management consultants, contractors and owners/clients were invaluable and their generosity is gratefully acknowledged.

REFERENCES

- Abbasi, O., Noorzai, E., Jafari, K.G. and Golabchi, M. (2020). Exploring the causes of delays in construction industry using a cause-and-effect diagram: Case study for Iran. *Journal of Architectural Engineering*, 26(3): 05020008. [https://doi.org/10.1061/\(ASCE\)AE.1943-5568.0000431](https://doi.org/10.1061/(ASCE)AE.1943-5568.0000431)
- Abusafiya, H.A.M. and Suliman, S.M.A. (2017). Causes and effects of cost overrun on construction project in Bahrain: Part I (ranking of cost overrun factors and risk mapping). *Modern Applied Science*, 11(7): 20. <https://doi.org/10.5539/mas.v11n7p20>
- Ahady, S., Gupta, S. and Malik, R.K. (2017). A critical review of the causes of cost overrun in construction industries in developing countries. *International Research Journal of Engineering and Technology*, 4(3): 2550–2558.
- Akhund, M.A., Khoso, A.R., Khan, J.S., Imad, H.U. and Memon, K.M. (2019). Prompting cost overrun factors during PCP in construction projects. *Indian Journal of Science and Technology*, 12(4). <https://doi.org/10.17485/ijst/2019/v12i4/140936>
- Alemu, S.K. (2022). Construction time prediction model for public building projects. *Engineering, Construction and Architectural Management*, 29(5): 2183–2206. <https://doi.org/10.1108/ECAM-11-2020-0975>
- Allahaim, F. and Liu, L. (2015). Causes of cost overruns on infrastructure projects in Saudi Arabia. *International Journal of Collaborative Enterprise*, 5(1/2): 32–57. <https://doi.org/10.1504/IJCEN.2015.073176>
- Amini, S., Rezvani, A. and Tabasi, M. (2021). Identification and ranking significant factors causing cost overrun in the design and construction process of buildings using the BWM method (case study: City of Mashhad). *Urban Management Studies*, 12(44): 67–83.
- Amini, S., Rezvani, A., Tabassi, M. and Sadati, S.S.M. (2023). Causes of cost overruns in building construction projects in Asian countries: Iran as a case study. *Engineering, Construction and Architectural Management*, 30(7): 2739–2766. <https://doi.org/10.1108/ECAM-05-2021-0445>
- Ammar, T., Abdel-Monem, M. and El-Dash, K. (2022). Risk factors causing cost overruns in road networks. *Ain Shams Engineering Journal*, 13(5): 101720. <https://doi.org/10.1016/j.asej.2022.101720>
- Ashtari, M.A., Ansari, R., Hassannayebi, E. and Jeong, J. (2022). Cost overrun risk assessment and prediction in construction projects: A Bayesian network classifier approach. *Buildings*, 12(10): 1660. <https://doi.org/10.3390/buildings12101660>
- Aung, T., Liana, S.R., Htet, A. and Bhaumik, A. (2023). Using machine learning to predict cost overruns in construction projects. *Journal of Technology Innovations and Energy*, 2(2): 1–7. <https://doi.org/10.56556/jtie.v2i2.511>

- Balaji, H.V. and Venugopal, P. (2020). Impact of cost overrun factors in constructing apartments in Tamil Nadu. *International Journal of Innovative Technology and Exploring Engineering*, 9(3): 456–459. <https://doi.org/10.35940/ijitee.C8092.019320>
- Balali, V., Zalavadia, A. and Heydarian, A. (2020). Real-time interaction and cost estimating within immersive virtual environments. *Journal of Construction Engineering and Management*, 146(2): 04019098. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0001752](https://doi.org/10.1061/(ASCE)CO.1943-7862.0001752)
- Balta, G.C.K., Dikmen, I. and Birgonul, M.T. (2021). Bayesian network-based decision support for predicting and mitigating delay risk in TBM tunnel projects. *Automation in Construction*, 129: 103819. <https://doi.org/10.1016/j.autcon.2021.103819>
- Bentil, E., Nana-Addy, R.E., Asare, E.K. and Fokuo-Kusi, A. (2017). The level of existence and impact of cost and time overruns of building construction projects in Ghana. *Civil and Environmental Research*, 9: 36–46.
- Bishnoi, S. and Hooda, B.K. (2022). Decision tree algorithms and their applicability in agriculture for classification. *Journal of Experimental Agriculture International*, 44(7): 20–27. <https://doi.org/10.9734/jeai/2022/v44i730833>
- Car-Pušić, D., Tijanić, K., Marović, I. and Mladen, M. (2020). Predicting buildings construction cost overruns on the basis of cost overruns structure. *Scientific Review: Engineering and Environmental Studies*, 29(3): 366–376. <https://doi.org/10.22630/PNIKS.2020.29.3.31>
- Creswell, J.W. and Clark, V.L.P. (2017). *Designing and Conducting Mixed Methods Research*. 3rd Ed. Thousand Oaks, CA: SAGE Publications.
- Dabirian, S., Khanzadi, M. and Abdollahi, M. (2020). Investigation of the most important factors affecting cost overruns in construction projects using hybrid system dynamic and DEMATEL method. *Sharif Journal of Civil Engineering*, 36(2): 13–28. <https://doi.org/10.24200/j30.2019.51915.2482>
- Derakhshanlavijeh, R. and Teixeira, J.M.C. (2017). Cost overrun in construction projects in developing countries, gas-oil industry of Iran as a case study. *Journal of Civil Engineering and Management*, 23(1): 125–136. <https://doi.org/10.3846/13923730.2014.992467>
- Durdyev, S., Ismail, S. and Abu Bakar, N. (2012). Factors causing cost overruns in construction of residential projects; case study of turkey. *International Journal of Science and Management*, 1(1): 3–12.
- El-Dash, K.M.M., Ramadan, O.M.O. and Youssef, W.M.M.A. (2019). Duration prediction models for construction projects in the Middle East. *Engineering, Technology and Applied Science Research*, 9(2): 3924–3932. <https://doi.org/10.48084/etasr.2531>
- Eltoukhy, M.G. and Nassar, A.H. (2021). Using support vector machine (SVM) for time and cost overrun analysis in construction projects in Egypt. *International Journal of Civil Engineering and Technology*, 12(3): 5–22. <https://doi.org/10.34218/IJCIET.12.3.2021.002>
- Esmaili, I. and Kashani, H. (2022). Managing cost risks in oil and gas construction projects: root causes of cost overruns. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 8(1): 04021072. <https://doi.org/10.1061/AJRUA6.0001193>
- Farzanegan, M.R. and Batmanghelidj, E. (2023). Understanding economic sanctions on Iran: A survey. *The Economists' Voice*, 20(2): 197–226. <https://doi.org/10.1515/ev-2023-0014>
- Flyvbjerg, B. and Bester, D.W. (2022). How (in)accurate is cost-benefit analysis? Data, explanations and suggestions for reform. In J.A. Gómez-Ibáñez and Z. Liu (eds.), *Infrastructure Economics and Policy: International Perspectives*. Cambridge, MA: Lincoln Institute of Land Policy, 174–196.

- Flyvbjerg, B., Skamris Holm, M.K. and Buhl, S.L. (2003). How common and how large are cost overruns in transport infrastructure projects? *Transport Reviews*, 23(1): 71–88. <https://doi.org/10.1080/01441640309904>
- Gebrehiwet, T. and Luo, H. (2017). Analysis of delay impact on construction project based on RII and correlation coefficient: Empirical study. *Procedia Engineering*, 196: 366–374. <https://doi.org/10.1016/j.proeng.2017.07.212>
- Ghazal, M.M. and Hammad, A. (2022). Application of knowledge discovery in database (KDD) techniques in cost overrun of construction projects. *International Journal of Construction Management*, 22(9): 1632–1646. <https://doi.org/10.1080/15623599.2020.1738205>
- Ghazizade, M.J., Koulivand, H., Safari, E. and Heidari, L. (2021). Petrochemical waste characterization and management at Pars Special Economic Energy Zone in the south of Iran. *Waste Management & Research: The Journal for a Sustainable Circular Economy*, 39(2): 199–208. <https://doi.org/10.1177/0734242X20922585>
- Gotlur, K., Pujar, G., Nayak, S. and Bh, J.R. (2020). Review on prediction of cost overrun in construction projects using non-traditional methods. *International Research Journal of Engineering and Technology*, 7(4): 5728–5733.
- Gupta, C. and Kumar, C. (2020). Study of factors causing cost and time overrun in construction projects. *International Journal of Engineering Research and Technology*, 9(10): 202–206.
- Heravi, G. and Mohammadian, M. (2021). Investigating cost overruns and delay in urban construction projects in Iran. *International Journal of Construction Management*, 21(9): 958–968. <https://doi.org/10.1080/15623599.2019.1601394>
- Ibrahim, A.K., Yasin, M.N.B. and Hassan, M.H. (2024). A Study on causative factors of cost overrun in construction projects in Mogadishu, Somalia. *Tropical Scientific Journal*, 3(1): 1–16.
- Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B. and Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622: 178–210. <https://doi.org/10.1016/j.ins.2022.11.139>
- Jadhav, D., Konnur, B.A. and Patil, S. (2020). Analysis of factors causing cost overruns in residential building construction projects. *International Journal of Research in Engineering, Science and Management*, 3(7): 308–311.
- Johari, Z., Cheraghi, M. and Ardakani, S.S. (2018). Environmental risk assessment of Ilam Petrochemical Company using analytical network analysis and the technique for order of preference by similarity to ideal solution methods in 2016. *Journal of Ilam University of Medical Sciences*, 26(5): 79–88. <https://doi.org/10.29252/sjimu.26.5.79>
- Johnson, R.M. and Babu, R.I.I. (2018). Time and cost overruns in the UAE construction industry: A critical analysis. *International Journal of Construction Management*, 20(5): 402–411. <https://doi.org/10.1080/15623599.2018.1484864>
- Kamaruddeen, A.M., Chang, F.S. and Wahi, W. (2020). A study on factors causing cost overrun of construction projects in Sarawak, Malaysia. *Civil Engineering and Architecture*, 8(3): 191–199. <https://doi.org/10.13189/cea.2020.080301>
- Keng, T.C., Mansor, N. and Ching, Y.K. (2018). An exploration of cost overrun in building construction projects. *Global Business and Management Research: An International Journal*, 10(3): 638–646.
- Laudati, D. and Pesaran, M.H. (2023). Identifying the effects of sanctions on the Iranian economy using newspaper coverage. *Journal of Applied Econometrics*, 38(3): 271–294. <https://doi.org/10.1002/jae.2947>

- Leu, S.-S., Liu, Y. and Wu, P.-L. (2023). Project cost overrun risk prediction using hidden Markov chain analysis. *Buildings*, 13(3): 667. <https://doi.org/10.3390/buildings13030667>
- Li, T., Shen, H., Yuan, Q., Zhang, X. and Zhang, L. (2017). Estimating ground-level PM_{2.5} by fusing satellite and station observations: A geo-intelligent deep learning approach. *Geophysical Research Letters*, 44(23): 11985–11993. <https://doi.org/10.1002/2017GL075710>
- Liravinia, B., Modiri, M. and Hafshjani, K.F. (2023). Identification and evaluation of the design risks in Iran's oil, gas and petrochemical industry. *Journal of Information and Organizational Sciences*, 47(1): 21–37. <https://doi.org/10.31341/jios.47.1.2>
- Mahamid, I. (2014). Contractors' perception of risk factors affecting cost overrun in building projects in Palestine. *IES Journal Part A: Civil and Structural Engineering*, 7(1): 38–50. <https://doi.org/10.1080/19373260.2013.854180>
- Mahdavian, A., Shojaei, A., Salem, M., Yuan, J.S. and Oloufa, A.A. (2021). Data-driven predictive modeling of highway construction cost items. *Journal of Construction Engineering and Management*, 147(3): 04020180. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0001991](https://doi.org/10.1061/(ASCE)CO.1943-7862.0001991)
- Mittal, M., Sharma, R.K. and Singh, V.P. (2019). Performance evaluation of threshold-based and k-means clustering algorithms using Iris dataset. *Recent Patents on Engineering*, 13(2): 131–135. <https://doi.org/10.2174/1872212112666180510153006>
- Moghayed, A. and Windapo, A. (2022). Modelling the uncertainty of cost and time in highway projects. *Infrastructure Asset Management*, 9(2): 73–88. <https://doi.org/10.1680/jinam.21.00004>
- Monteiro, F.P., Sousa, V., Meireles, I. and Oliveira Cruz, C. (2021). Cost modeling from the contractor perspective: Application to residential and office buildings. *Buildings*, 11(11): 529. <https://doi.org/10.3390/buildings11110529>
- Mrad, C., Titirla, M.D. and Larbi, W. (2021). Comparison of strengthening solutions with optimized passive energy dissipation systems in symmetric buildings. *Applied Sciences*, 11(21): 10103. <https://doi.org/10.3390/app11210103>
- Nassar, Y.S. and Erzajj, K.R. (2022). Developing models to predict the effect of crises on construction projects using MLR technique. *Periodicals of Engineering and Natural Sciences*, 10(2): 455–466. <https://doi.org/10.21533/pen.v10i2.2769>
- Nyimbili, F. and Nyimbili, L. (2024). Types of purposive sampling techniques with their examples and application in qualitative research studies. *British Journal of Multidisciplinary and Advanced Studies*, 5(1): 90–99. <https://doi.org/10.37745/bjmas.2022.0419>
- Omran, A., Saleh, M.S.H. and Gebril, A.W.O. (2023). Factors causing time and cost overruns of construction projects in Malaysia. *Design, Construction, Maintenance*, 3: 179–186. <https://doi.org/10.37394/232022.2023.3.15>
- Pencina, M.J., D'Agostino, R.B. and Massaro, J.M. (2013). Understanding increments in model performance metrics. *Lifetime Data Analysis*, 19(2): 202–218. <https://doi.org/10.1007/s10985-012-9238-0>
- Petrusheva, S., Car-Pušić, D. and Zileska-Pancovska, V. (2019). Support vector machine based hybrid model for prediction of road structures construction costs. *IOP Conference Series: Earth and Environmental Science*, 222: 012010. <https://doi.org/10.1088/1755-1315/222/1/012010>
- Piryonesi, S.M. and El-Diraby, T.E. (2019). Data analytics in asset management: Cost-effective prediction of the pavement condition index. *Journal of Infrastructure Systems*, 26(1): 04019036.
- Plebankiewicz, E. and Wieczorek, D. (2020). Prediction of cost overrun risk in construction projects. *Sustainability*, 12(22): 9341. <https://doi.org/10.3390/su12229341>

- Project Management Institute (2017). *A Guide to the Project Management Body of Knowledge (PMBOK® Guide)*. 6th Ed. Newtown Square, PA: Project Management Institute.
- Ragheb, A.M., Massoud, H.H. and Hamisa, M.T. (2020). Developing an ANN model to predict the contractor's actual cost overruns for lump sum contract projects in Egypt. *Al-Azhar University Civil Engineering Research Magazine*, 42(3): 98–126.
- Rajput, B. and Agarwal, A. (2020). Selection of construction equipment supplier using analytical hierarchy process. *IUP Journal of Management Research*, 19(3): 7–15.
- Sardroud, A.M., Kameli, M., Hosseinalipour, M., Ahmed, S.M. and Behruman, M. (2021). Improving maintenance performance by developing an IFC BIM/RFID-based computer system. *Journal of Ambient Intelligence and Humanized Computing*, 12: 3055–3074.
- Senouci, A., Ismail, A. and Eldin, N. (2016). Time delay and cost overrun in Qatari public construction projects. *Procedia Engineering*, 164: 368–375. <https://doi.org/10.1016/j.proeng.2016.11.632>
- Shahidi-Nia, O. (2022). PETZONE petrochemical output seen to hit 19 m tons by late Mar. *Tehran Times*, 5 March.
- Sharma, V., Zaki, M., Jha, K.N. and Krishnan, N.M.A. (2022). Machine learning-aided cost prediction and optimization in construction operations. *Engineering, Construction and Architectural Management*, 29(3): 1241–1257. <https://doi.org/10.1108/ECAM-10-2020-0778>
- Shiha, A., Dorra, E.M. and Nassar, K. (2020). Neural networks model for prediction of construction material prices in Egypt using macroeconomic indicators. *Journal of Construction Engineering and Management*, 146(3): 04020010. [https://doi.org/10.1061/\(ASCE\)CO.1943-7862.0001785](https://doi.org/10.1061/(ASCE)CO.1943-7862.0001785)
- Shoar, S., Chileshe, N. and 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. <https://doi.org/10.1016/j.jobe.2022.104102>
- Sohu, S., Ansari, A.A. and Jhatial, A.A. (2021). Most common factors causing cost overrun with its mitigation measure for Pakistan construction industry. *International Journal of Sustainable Construction Engineering and Technology*, 11(2): 256–261. <https://doi.org/10.30880/ijscet.2021.11.02.032>
- Srinivasan, N.P., Gowtham, T., Kumar, A.R.R., Baalaji, M. and Kumar, S.N. (2021). A review on feasible cost prediction model in construction projects. *Information Technology in Industry*, 9(3): 537–543.
- Tajziyehchi, N., Moshirpour, M., Jergeas, G. and Sadeghpour, F. (2020). A predictive model of cost growth in construction projects using feature selection. In *2020 IEEE Third International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. New York: IEEE, 142–147. <https://doi.org/10.1109/AIKE48582.2020.00029>
- Tayyab, M., Furkhan, M., Rizwan, M., Jameel, M. and Chadee, A. (2023). A study on factors influencing cost overrun in high-rise building construction across India. *Journal of Smart Buildings and Construction Technology*, 5(1): 52–83. <https://doi.org/10.30564/jsbct.v5i1.5489>
- Tijanć, K., Car-Puščić, D. and Šperac, M. (2019). Cost estimation in road construction using artificial neural network. *Neural Computing and Applications*, 32: 9343–9355.
- Xu, J. (2023). Systematic analysis and application prospect of decision tree. *Highlights in Science, Engineering and Technology*, 71: 163–170. <https://doi.org/10.54097/hset.v71i.12687>