buildings

MDPI

Article

Risk Identification, Assessments, and Prediction for Mega Construction Projects: A Risk Prediction Paradigm Based on Cross Analytical-Machine Learning Model

Debalina Banerjee Chattapadhyay [1], Jagadeesh Putta [1,*] and Rama Mohan Rao P [2]

[1] School of Civil Engineering, Vellore Institute of Technology, Vellore 632014, India; banerjee.chattapadhyay2015@vit.ac.in
[2] Centre for Disaster Mitigation and Management, Vellore Institute of Technology, Vellore 632014, India; rao_pannem@vit.ac.in
* Correspondence: p.jagadeesh@vit.ac.in; Tel.: +91-94447-12064

Abstract: Risk identification and management are the two most important parts of construction project management. Better risk management can help in determining the future consequences, but identifying possible risk factors has a direct and indirect impact on the risk management process. In this paper, a risk prediction system based on a cross analytical-machine learning model was developed for construction megaprojects. A total of 63 risk factors pertaining to the cost, time, quality, and scope of the megaproject and primary data were collected from industry experts on a five-point Likert scale. The obtained sample was further processed statistically to generate a significantly large set of features to perform K-means clustering based on high-risk factor and allied sub-risk component identification. Descriptive analysis, followed by the synthetic minority over-sampling technique (SMOTE) and the Wilcoxon rank-sum test was performed to retain the most significant features pertaining to cost, time, quality, and scope. Eventually, unlike classical K-means clustering, a genetic-algorithm-based K-means clustering algorithm (GA–K-means) was applied with dual-objective functions to segment high-risk factors and allied sub-risk components. The proposed model identified different high-risk factors and sub-risk factors, which cumulatively can impact overall performance. Thus, identifying these high-risk factors and corresponding sub-risk components can help stakeholders in achieving project success.

Keywords: risk management; risk identification; construction megaproject; quantitative; machine learning; evolutionary computing

Citation: Banerjee Chattapadhyay, D.; Putta, J.; Rao P, R.M. Risk Identification, Assessments, and Prediction for Mega Construction Projects: A Risk Prediction Paradigm Based on Cross Analytical-Machine Learning Model. *Buildings* **2021**, *11*, 172. https://doi.org/10.3390/buildings11040172

Academic Editors: António Aguiar Costa and Manuel Parente

Received: 8 March 2021
Accepted: 14 April 2021
Published: 17 April 2021

1. Introduction and Background

The exponential rise in global competitiveness has triggered every economy to strengthen its socioeconomic aspects in terms of better social conditions, economic strengths, technological prospects, and global recognition. This has motivated economies to improve their intrinsic conditions, including infrastructures. This has resulted in the undertaking and execution of large complex projects or megaprojects involving substantial funding and spanning over a long duration of time. Typically, a megaproject can be characterized in terms of its complexity, substantial investment [1–3] uncertainty, dynamism, dynamic interfaces, vital sociopolitical and external influences, and large construction time [4,5]. Among the different kinds of projects, megaprojects are often considered the most complex and uncertain due to multiple stakeholder dependencies, substantial investment, greater community involvement, a complex decision-making environment [2], and the likelihood of risk [5]. Moreover, the unpredictable nature of megaprojects, even with precalculated and calibrated project details, signifies a proneness to risk. Undeniably, the failure of a megaproject often results in the collapse of funding agents, substantial financial losses, and project uphold for the long term, sometimes forever [6]. As theses are projects involving multibillion investments

and has a probability of undergoing various risk forces, there is a need to inculcate optimal project management policies and execution control [6–9].

Identifying inherent risk and its timely avoidance has always been a challenge, especially with megaprojects. Construction megaprojects, being highly dependent on local conditions, social acceptance, and government affirmation, undergo political risk that prolongs project completion time and imposes substantial financial risk [10]. The lack of timely funds and resources also adversely impacts the project. However, the severity and complexity of these risk factors play different roles in different megaprojects. These risk factors can more severely impact some projects, while they may have less impact on others. For instance, a megaproject completely funded by private players has lower political risks, provided it gains initial approval with certain standard terms and agreements. On the contrary, a government-funded megaproject, especially one related to construction, undergoes significant turbulence due to political regime changes and local conditions [9]. To alleviate such risks in construction megaprojects, identifying risks and distributing them across stakeholders can have an affirmative impact on project success.

Megaprojects also require an assessment of inter-association among the risks to segment them into proper categories. This can help in efficient risk management and handling. The characteristics of megaprojects and associated risk probability broaden the horizon for academic industries to assess the different cases and segment the key possible risks and alleviating (optimal) measures to ensure proactive decision making across the project cycle [11].

Since risks and their severity vary from one project to another based on their nature, risk distribution, stakeholders, or investment assessing the risk for each kind of project can be of great significance [11]. In order to alleviate any possible detrimental losses and enduring problems in construction megaprojects, assessing the different risk factors is especially important [12]. Identifying different risks, their inter-relationship, and eventual driving forces, along with corresponding alleviation measures, can be of utmost significance. It can help project management and allied decision-making environments become more specific and favorable, especially for a megaproject [7,12].

In sync with contemporary construction megaprojects, the activities pertaining to project planning, risk identification, qualitative risk analysis, quantitative risk analysis, risk response planning, risk monitoring, and control are necessary [13–17]. Identifying, assessing, and responding to potential risk can help construction megaprojects reduce detrimental effects over project cycles [9,18,19]; however, being a multistakeholder, multiagent investment, or a locally sociopolitical-support-driven project, assessing their potential risk associations is vital [14,16–18]. It can avoid detrimental effects and help in achieving better rewards [13,18,20]. Interestingly, although many studies address risk management in small-scale or medium-sized projects [21], very few have considered risk assessment in construction megaprojects [11,22–26]. This broadens the horizon for researchers to assess the risk management of megaprojects, so as to derive a holistic model that addresses the different potential risks in (construction) megaprojects [9,18,19]. In the last few decades, the high-pace rises in the number and value of megaprojects [27] and the heterogeneity of different risk factors have been witnessed in varied or distinct megaprojects; therefore, the same risk management model or framework cannot be adopted for all megaprojects [26]. This is because, as the size, complexity, and multiparty involvement increases, the proneness to risk increases in parallel [19,25,28]. Therefore, assessing risks for construction megaprojects as a distinct case is a must [26]. This, as a result, can help in achieving proper risk identification, risk severity assessment, inter-risk association, and risk distribution among the agents or stakeholders to avoid delayed projects and financial losses [9,13,26].

As stated, in major existing studies, a qualitative assessment has been made to identify risk factors and probable avoidance measures. Ironically, the existing methods do not contribute to a project-specific risk management framework and hence cannot be adopted in all megaprojects [13]. On the contrary, despite the high-pace rise in the construction sector across global economies and the allied likelihood of risks, very few studies address

risk management in construction megaprojects [25,29]. Interestingly, the at-hand studies identified execution risk, economic risk, political risk, environmental risk, technological risk, and social risks as key risk factors impacting construction megaprojects [26,30]. Despite the analytical-approach-based identification, these studies could not identify relationships and near associations among other risk factors such that risk management and decisions could be improved to accomplish target endeavors [26]. In addition to the above-stated key risk factors (execution, economic, political, environmental, technological, and social), there can be other inter-related risks that have a cumulative impact on the project cycle if they remain unaddressed. Therefore, clustering different risk factors based on their association closeness can help in segmenting key risks and (closing) sub-risk factors. This, as a result, can enable better risk distribution and handling to ensure a project's success.

To identify the different risk factors and allied sub-risk components, exploiting inter-risk connectivity can be of utmost significance. Merely identifying risk factors based on certain qualitative or quantitative methods cannot generalize a risk management framework. Undeniably, quantitative methods or analytical methods with expert responses toward risk factors can help researchers or industries to identify major (broad) risks; however, it might fail in addressing sub-risks, which can have a cumulative impact on the overall project success.

The main aim of the study was to design a dynamic framework for the assessment of the complexities of risks in megaproject development. With this aim of the study, a new knowledge-driven, machine-learning-based risk identification framework is developed for construction megaprojects. Since it is a cross model, we exploit both a quantitative research method to collect and process experts' perceptions and suggestions toward different risk factors in construction megaprojects, as well as machine learning algorithms to cluster the different risk factors based on their dependencies and cumulative impact on project performance.

It can help in reducing any detrimental effects on the life cycle of a construction megaproject. Key contributions in this research paper can be summarized as follows:

1. Unlike classical qualitative- or even limited quantitative-study-based risk identification, in this study, we considered industry-expert-knowledge-driven risk identification and verification. In other words, we obtained responses from experts handling (or having handled) construction megaprojects toward different potential risk factors and their impact on the project lifecycle;
2. Due to the study being a knowledge-driven risk management approach, we performed a statistical assessment of the different risk factors and their impact on the project lifecycle and performance. Moreover, deriving a cumulative risk score from the major risk factors, machine learning was applied to segment other correlated sub-risk factors for better decision making;
3. Since the study used a multiple risk scenario, we processed the statistical output of the different risks and their severity using a GA-based K-means clustering algorithm. The proposed GA K-means algorithm exploits the respective weights of major risks as well as sub-risk factors to segment (or group) overall risks into broad categories;
4. The identification of the major risks along with the closest sub-risks can enable better risk management in terms of risk awareness, risk distribution, proactive decisions, and execution. Cumulatively, it can be of vital significance to avoid detrimental losses that could be caused due to hidden risks or inferiorly weighted risks in construction megaprojects;
5. Since the proposed research contributes a risk identification framework based on expert-driven knowledge and machine-learning-based predicted outcomes, its efficacy toward project risk management is more justifiable and adoptable.

2. Literature Review

2.1. Characteristics of the Megaproject

Typically, key factors characterizing a project as a megaproject include an investment over USD 1 billion, high dynamism, feasible intangible benefits, and striking long-term results and allied benefits [1,3]. Megaprojects are also characterized in terms of significantly high complexity and design risk and therefore undergo high engineering and design risks throughout the project development [18]. Some of the well-known megaprojects in the world are Taipei 101 tower in Taipei, Taiwan (tallest building in the world, 508 m), Roosevelt Dam Bridge in Arizona USA, (longest steel arch bridge), Big I reconstruction project in New Mexico, USA, (largest freeway project in the state) [23], characterizing high complexity in design, dynamism, huge budget, volatility, and uncertainty. This imposes a proneness to risks that require identification and proactive management for better project performance [31]. This may be facilitated by complexity measurement [31] and monitoring project performance at preconstruction and construction phases. Some of the tools at hand for performance measurement systems include cost/schedule-based systems called earned value management systems (EVMSs), the balanced score card (BSC), and key performance indicators [32]. Risk, on the other hand, is often measured by risk performance index based on project performance tools [32]. Researchers at North Carolina State University conducted a study to model the payout curve patterns for completed design and build megaprojects using macro/micro approach during preconstruction phase and model approach during construction phase [33]. The macro approach creates project payout models based on data from past projects, while the micro approach builds the project's payout curve based on the anticipated cost of the various construction activities involved. The model during the construction phase is based on the estimated expenditure forecasts by the contractors. The outcomes from the study also revealed the "key factors" influencing megaprojects such as let-date delay, merger process, scope creep, consent and approval for railroad, permits, utility relocation during the preconstruction phase and utilities, railroads, Right of Way (ROW), and permits during the construction phase [33].

In sync with megaprojects, risk can be defined as any unexpected event that degrades the project's goals, and as a result, the project might fail to serve the expected endeavors of the allied stakeholders [34]. It is also defined as "an uncertain event that, in case [it] occurs, can have affirmative as well as negative effect on a project's goal and endeavours" [15]. Each risk factor often has certain visible affirmative and negative consequences. Risk is an especially important factor that needs to be considered, as it can influence both the cost–benefit analysis, local perceptions, construction costs, execution time, and major financial variables [12].

The impact of risk relies on the causative event and on the way the responsible management addresses it and deals with it. A typical construction megaproject possesses numerous risk factors that can cause prolonged construction delays, significantly substantial financial losses over the project's life cycle, and stakeholders' perceptual differences. It results in project failure, and therefore, certain optimistic and specific management actions, including risk assessment and mitigation measures, need to be taken [6].

2.2. Risk Identification, Assessment, and Management: Current Scenario

Different researchers [35–38] have stated that the majority of probable risks pertaining to the megaproject, and their root cause can be identified by assessing the organizational dynamics and multidisciplinary characteristics of today's business environment [39–41]. Due to megaprojects being a multi-stakeholder-based business environment, the proactive participation of the varied autonomous processing elements, technologies, and stakeholders in the extent or likelihood of uncertainty and risk distribution turns out to be broadened over a wide area of enterprise and allied partners [39–41]. Under these circumstances, managing megaprojects requires going beyond a simple analysis of the cost and dates and understanding the cause of any uncertainty [42].

Many studies have revealed that the predominant risks involved in megaprojects are political risks, design risks, economic risks [2]. Among the major risks, political risks represent uncertain financing that often results in potential revenue declination. Being unpredicted in nature, catastrophic loss also affects megaprojects significantly in terms of delay and financial losses. In reference to construction megaprojects, in addition to the above-stated risk factors, environmental risks, execution risks, and social risks also have a decisive impact on a project's success and target endeavors. Notably, among the major risks, those that can have a direct impact on a project's execution, performance, and delay turn out to be the most critical and hence require early identification and resolution [6,43].

Observing the above-stated risk classifications, it can easily be inferred that there is no broad and homogenous classification approach in which authors can exploit both interdependencies of the risk factors and their impact on overall project success. The at-hand risk classification methods are undeniably confined because of their inability to identify all types of risks with a certain probability over a mega construction project, especially based on public and private stakeholders' joint ventures. Moreover, most of the existing risk identification or classification methods lack the source-oriented groupings, expert consensus, and inter-risk dependency needed for precise risk exposure pertaining to a mega construction project [44].

Although each project requires a multidimensional analysis to assess success probability, inherent risks, and viable solutions, it is especially important for megaprojects because of their broader complexity and substantial investment. Harvet [45] stated that a significantly large number of projects often fail due to the increases in project complexity over time. This raises the question as to whether the at-hand industrial risk management policies and allied standards are effective in avoiding uncertain losses and failure [44]. Additionally, risk management and allied practices are not the same in all projects, as risks do not impact all projects in the same manner or to a similar extent [38].

Irrespective of the project size, the risk management process requires identifying inherent risks and the project's optimal avoidance measures [42]. More specifically, the megaproject risk assessment and management practices involve identifying the optimal strategies to reduce risk probability, comprising the way the possible risks are shared among the stakeholders and the risks that could be transferred [6].

In reference to a megaproject, the key parameters of time, cost, and scope constitute classical performance measurements, often collectively called the "triple constraint" or the "iron triangle," which need to be addressed, and project management practices need to be performed to retain target endeavors [46]. Based on an analytical assessment method, lower performance can be characterized in terms of multiple factors pertaining to a megaproject, such as resource constraints, higher complexity, the lack of realism in estimates, inefficient management, and public (stakeholders) resistance due to local causes and/or politically motivated social agitation [6,47–49]. Additionally, there are also some hidden risk factors, which are minimally addressed in the literature, especially in reference to megaprojects. These are unfair bids, biased leasing, and unrealistic or undervalued bids. These factors often push a megaproject to undergo financial losses and a default-driven delay and often lead toward the risk of negotiation that, in general, turns out to cause delays for a construction megaproject [22,47].

2.3. Significance of the Present Study

Even though, in the last few years, a significantly large number of studies have been conducted on risk management practices in projects, very little research addresses risk assessment and allied management policies for megaprojects [22]. On the contrary, the need to assess risks pertaining to megaprojects has always been a key demand from industries [11,23], as it guides and sets up standards for management to make optimal and calibrated decisions in project planning and execution while retaining stakeholders' endeavors. Risk identification and a corresponding avoidance measure formulation are especially important in a megaproject, especially when a megaproject undertakes both

public and private agencies [30]. Although due to budgetary constraints, a few researchers have contributed models toward public–private partnerships in megaprojects, they have been primarily confined to financial management and resource optimization [30]. The exceedingly high uncertainty and risk impact of construction megaprojects make it especially important to have optimal risk management policies [23,50]. Certain authors have found that consideration of key factors such as "the imprecision, vagueness, and fuzziness of the risk factors is [especially important] for a construction project to appropriately deal with a contractor's project risks by using Fuzzy Set Theory (FST)"; however, the lack of specific methodologies, databases, and journals analyzed confine this generalization [14]. Taroun (2014) studied different risk models and measures in construction projects [51]. Assessing the different definitions, risk elements, and allied concepts of risk models in construction projects, the authors identified key tools and theories and stated that there is a lack of a general framework to assess the risks and their corresponding impact on construction projects [19,50,51]. However, these authors could not address concerns toward construction megaprojects, which have greater risk impacts and fluctuations than typical, smaller projects.

To classify different risk factors in construction projects, different models have been proposed. A generic classification model [52] was used, but it was aimed at classifying different risks in the early phase of a construction project cycle. The authors identified a few key risk factors such as cost risk, demand risk, financial market risk, and political risk. Performing a more in-depth assessment, the authors classified the different risk factors as follows: (1) cost risk: construction, maintenance, and operation; (2) demand risk: traffic forecasts and revenues; (3) financial market risk: future interest rates; (4) political risk: regulation, parallel public investment, and pricing in adjacent parts of the network. However, these risk factors were not found to have generalizable significance or impacts over the entire project cycle. Moreover, it failed to address any associations or dependencies between one risk factor and another that might have a decisive impact on overall project (megaproject) success.

A similar assessment was carried out by Little [53], who classified construction project risk as political risk, construction (or execution) risk, operation and maintenance risk, legal and contractual risk, income risk, and financial risk. Bing et al. [54], on the other hand, classified risks based on their severity, considering macro-, meso-, and microlevel risks. Here, macro risks included exogenous risks, while endogenous risks were considered mesolevel risks. On the contrary, the microlevel signifies risks caused by stakeholder relationships constituted during the procurement process because of an inherent disparity between the public and private sectors in contract management. Rolstadas and Johansen [55], Krane et al. [56], and Westney and Dodson [57] differentiated in terms of contextual risk, in which the former stated that there is a prospective impact on outcomes or capital due to adverse business decisions, ineffective execution, and implementation decisions, or a lack of responsiveness toward organizational changes.

Recently, Krane et al. [56] and Krane et al. [58] distinguished risk factors based on the project's goals. They classified risk as operational risks (mainly pertaining to the project's operational goals, confined to the direct outcomes of the project), short-term strategic risks, and long-term strategic risks. Turner [59] differentiated risks as business risks and insurable risks (because of the unexpected set of events during a project cycle).

3. Materials and Methods

In principle, there are many research methods needed to fulfill various research needs [60]. The irony is that, while there are indeed many research methods, there is no perfect option. Nevertheless, some methods are better suited for tackling specific issues than others. This research is based on a mixed research paradigm involving both qualitative and quantitative approaches. Figure 1 illustrates the overall flow of the proposed framework for this study.

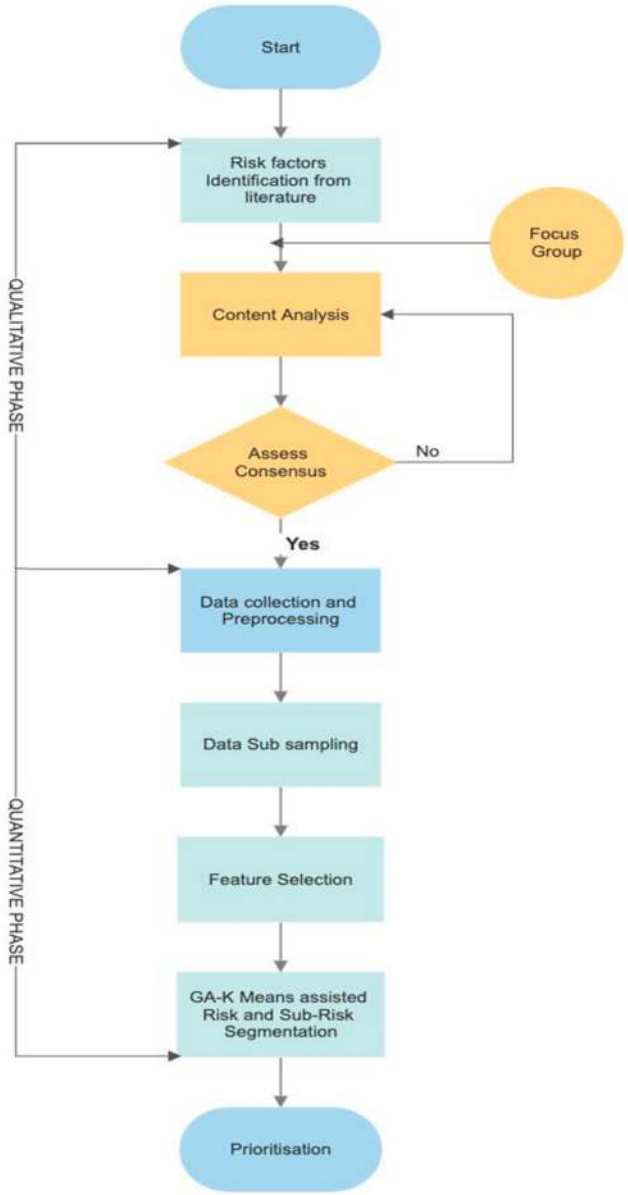

Figure 1. Process flow chart for the proposed methodology of research.

3.1. Qualitative Phase

As a part of the qualitative phase, a content analysis was performed with the help of a focus group based on the review of literature from articles, journals, textbooks, and other internet sources. A theme-based categorization of broad risk factors was performed, followed by coding and finding the various risks under each category. In total, 10 members were identified as a part of the focus group with 15–25 years of experience, ranging from industry experts to academicians in the area of construction management practice. Experts from the focus group were given two rounds of questions. The first round had two open-ended questions seeking their opinion on what major categories of risks megaprojects may be subjected to, while the second round had 63 different risk factors identified from the literature and sought their agreement. A content analysis [61] was conducted from the transcripts of the interview with the help of the codes used. The codes depicting the greatest number of agreements were considered for further analysis. The consensus was supposed to have been achieved when 80% of the experts agreed that the 63 risk factors were suitable for further quantitative analysis. This could eventually help in arriving at the different risk factors and allied sub-risk components and also inter-risk connectivity that

could affect project performance. Thus, by conducting the interviews, it was possible to consider the relevance of the risk factors identified in the literature.

3.2. Quantitative Phase

The instrument for the quantitative phase comprised a questionnaire (mailed and online). The approach was considered necessary because it provides stronger empirical research evidence for explaining phenomena that will enable the researcher to address "how much" and "how many." Thus, after successfully identifying the different risk factors pertaining to the aspects of time, cost, quality, and scope, the experts were asked to state their views toward the impact of the different risk factors on project success and endeavors, using a scale from 1 to 5. The analysis was based on the inputs from the respondents on a five-point scale in which the respondents were asked to quantify the impact of each risk factor on the project objective of time cost quality and scope. Responses with an impact of 1 signify disagreement (interpreted as extremely low impact), while that with an impact of 5 indicate strong agreement (extremely high impact). The linguistic definition of the impact of risk factors was interpreted as shown below (Table 1).

Table 1. Linguistic definition of impact of risk factors adapted from Andric et al [62].

Impact	Rating
Extremely Low	1
Low	2
Moderate	3
High	4
Extremely High	5

A total of 150 questionnaires (Appendix A) were distributed, and among these, 91 completed questionnaires were successfully received. This represented a response rate of above 60%. From the responses received, 70 were considered to be valid and were used for the analysis part. Among the 70 valid responses in this research, 5 (7.14%) responses were from individual consultants, 40 (57.14%) were from Engineers, 4 (5.71%) were from Project Managers, 5 (7.14%) were from contractors or vendors, and 16 (22.86%) project team members. Most of the respondents had worked on the project for between 5 and 15 years. First, the responses from different experts toward varied contemporary risk factors, and their impact on project endeavors or success were obtained and processed for cumulative averaging. Thus, for obtaining the average rank for the different risk factors and their impact on a project's endeavors (success), a new heuristic-based, K-means clustering algorithm that clusters different risk components into optimal groups, containing different sub-risk factors to be handled cumulatively, was designed. Notably, unlike classical K-means-based clustering, in this research, a GA algorithm was applied to optimize the centroid estimation of the K-means algorithm, which identifies the central risk factors with a decisive impact and the closely connected sub-risk factors having an undeniable impact on project performance. This, as a result, provides identification of broad or major risk factors with corresponding sub-risk factors, which can help the management framework to distribute the risk factors optimally among the multiple stakeholders.

3.3. Research Questions

Considering the overall research goals and allied methodological paradigms, some research questions have been formulated. These questions assess whether the proposed methodology can accomplish overall research goals or not. The key questions defined are given as follows:

- Research Question 1 (RQ1): Are the risk factors pertaining to small- and midsize projects different from those of construction megaprojects?
- Research Question 2 (RQ2): What are the key risk and sub-risk components impacting construction megaprojects with respect to cost, time, quality, and scope?

- Research Question 3 (RQ3): Can the use of quantitatively or empirically driven risk factors and their severity assessment help in conceptualizing a risk management framework with risk–sub-risk segmentation for better multistakeholder megaproject (risk) management?
- Research Question 4 (RQ4): Can descriptive analysis of expert consensus toward the risk severity feature with the Wilcoxon rank-sum test as a feature selection method and a multi-objective-based GA–K-means algorithm yield optimal risk identification in construction megaprojects?
- Research Question 5 (RQ5): Can the use of expert-knowledge-driven, heuristic-assisted K-means clustering enable the segmenting of key risks and allied sub-risk factors for better risk management distribution in construction megaprojects?

3.4. System Model

As indicated in the previous sections, the proposed model encompasses both quantitative (say, empirical) analysis, followed by machine-learning-based risk segmentation, so as to identify the most critical risk factors along with sub-risk components. Thus, the overall method consists of the following steps:

1. Data Collection and Statistical Analysis;
2. Data Sub-Sampling;
3. Feature Selection;
4. GA–K-Means-Assisted Risk and Sub-Risk Segmentation.

Details of the above-stated research methods are given in the subsequent section.

3.4.1. Data Collection and Preprocessing

This phase can be stated as the initial and fundamental stage toward targeted risk and sub-risk segmentation (RSRS). Since there are no standard data available so far toward risk assessment and prediction in construction megaproject(s), we performed the quantitative method to collect different risk components and their impact on construction megaproject(s) under the current socioeconomic and political dynamism. Additionally, to ensure the reliability of the risk identified and its relevance toward contemporary megaproject conditions, we applied semi-structured interviews with industrial experts who have already handled construction megaprojects or are still working on one. Due to the study being a primary-data-based approach, we identified a total of 63 risk variables pertaining to the four key project aspects of cost, time, quality, and scope. These key risk factors, obtained and agreed upon during the qualitative phase, are presented in Table 2.

Table 2. Different risk factors found to have an impact on project performance [63].

Serial Number	Risk Categories	Risk Variables
1	Execution Risk Factors	Utility diversion
2		Inappropriate equipment and material quality
3		Permits and licenses
4		Poor equipment performance
5		Machinery failure/breakdown
6		Unforeseen site conditions
7		Incorrect take off calculation
8		Delayed supply of material and equipment
9		Delay in obtaining working drawings/reports/designs
10		Low skilled/incompetent workforce
11		Unavailability of materials, equipment, and labor.
12		Delay in obtaining temporary traffic regulation orders
13	Construction Risk Factors	Poor site coordination/work organization
14		Construction failure
15		Land acquisition for ROW
16		Inadequate preliminary survey and site information
17		Unrecognized soil structure/unforeseen ground condition
18		Delay in transport of ready-mix concrete (RMC)
19		Construction and implementation error from faulty design

Table 2. *Cont.*

Serial Number	Risk Categories	Risk Variables
20		Changes in material during construction
21		Deviations between specification and implementation
22		Supply chain breakdown/improper equipment and material quality
23		Site inaccessibility
24		Lack of site security for personnel and asset
25	Technical Risk Factors	Incompetency of designers
26		Design changes
27		Inadequate design and design errors
28		Modification to drawing/design
29		Unforeseen multiple modifications to project scope
30		Delay in obtaining preliminary drawings/reports
31		Revision in design standard
32		Inadequate project complexity analysis
33	Economic and Financial Risk Factors	Inflation
34		Foreign exchange rate and interest rate fluctuation
35		Changes in market conditions
36		Changes in taxes
37		Incorrect cost estimate
38		Financial difficulties/failure of subcontractor
39	Environmental Risk Factors	Natural Disaster
40		Adverse weather condition
41		Pollution and vibration
42		Geology, soil, and topography
43		Drainage pattern
44		Inadequate environmental analysis
45		Land cover (grass, asphalt, trees, water bodies)
46		Presence of quarries and mines
47	Social Risk Factors	Demands of local people
48		Public objections
49		Social issues (tree cutting, shrine removal)
50		Cultural and heritage sights
51		New stakeholders with changed request
52		Damage to property and persons
53		Multilevel decision-making bodies
54	Political Risk Factors	Changing government regulations/funding policy
55		Lack of moderators
56		Legal disputes
57		Political instability
58		Changes in local laws and standards (tax imposition)
59		Lack of political support
60		Political indecision
61		Change in government
62		Multilevel decision-making by government bodies for consent and approvals
63		Government intervention

Descriptive Statistics

We obtained experts' responses on a five-point scale, in which the above-stated risk variables (Table 1) were considered independent, while the performance (cost, quality, time, and scope) or the success parameters of the construction megaproject was considered as the common dependent variable. A total of 70 samples were collected from the respondents, thus obtaining 70 (expert) responses for 63 risk factors. We prepared a data table with average (using a mean statistical tool) values. The statistical analysis estimated the level of significance of each risk variable on the aforementioned project performance objectives. To enhance the data suitability and reliability further toward optimal risk segmentation, we performed descriptive analysis over the collected primary dataset or sample. We obtained 13 statistical parameters, as follows:

1. Minimum;
2. Maximum;
3. Mean;
4. Median;

5. Standard deviation;
6. Variance;
7. Percentile (First quartile, below 25%);
8. Percentile (Third quartile, below 75%);
9. Sum of responses with an impact of 1;
10. Sum of responses with an impact of 2;
11. Sum of responses with an impact of 3;
12. Sum of responses with an impact of 4;
13. Sum of responses with an impact of 5.

By finding the above-stated statistical parameters, an objective function has been obtained that signifies a threshold value characterizing significance or decisive impact or relation of each risk factor toward the stated four constructs of time, cost, quality, and scope. Since the minimum consensus of experts (i.e., the average of responses with an impact of 1 on the five-point scale) or the different consensus toward the varied risk variables can also have decisive significance in terms of characterizing contemporary risk factors in construction megaprojects, in addition to the classical statistical (descriptive) tools, we obtained the sums of responses with impacts of 1, 2, 3, 4 and 5, distinctly. Here, the sum of responses with an impact of 1 signifies disagreement (interpreted as having an extremely low impact), while that with an impact of 5 indicates strong agreement (interpreted as having a very high impact). Risk factors with 5 as the impact value were considered high-risk components (HRCs). Once obtaining the HRCs, we estimated the Euclidean distance between the HRCs and other risk parameters (Table 3). Thus, the risk factors near an HRC were obtained for each risk type (i.e., cost, time, quality, and scope) across the 63 risk factors given in Table 2. To find an optimal cumulative critical risk value (CRV) that has cumulative significance toward the four project aspects, we derived Equation (1).

$$\text{Optimal CRV}_i = \sqrt{(\text{COST}_i)^2 + (\text{TIME}_i)^2 + (\text{QUALITY}_i)^2 + (\text{SCOPE}_i)^2} \quad (1)$$

Table 3. Numbers of clusters vs. the silhouette coefficient with the standard K-means clustering algorithm for risk identification.

No. of Clusters	Silhouette Coefficient (K-Means)			
	Euclidian	**Manhattan**	**City-Block**	**Minkowski**
2	0.479	0.473	0.489	0.489
3	0.518	0.478	0.538	0.551
4	0.589	0.544	0.499	0.528
5	0.590 *	0.524	0.438	0.421
6	0.514	0.543	0.484	0.488
7	0.542	0.559 *	0.551	0.461
8	0.580	0.550	0.555 *	0.563 *
9	0.540	0.551	0.552	0.552
10	0.517	0.519	0.511	0.511

* Highest values.

As indicated in (1), we obtained the optimal CRV values Optimal CRV_i for each risk factor (where $i = 1, \ldots, 63$).

3.4.2. SMOTE Sub-Sampling

Obtaining the optimal CRV values, we synthesized a total of 10,000 samples using the synthetic minority over-sampling technique (SMOTE) [64]. This proposed a random sampling method that first obtains 10,000 samples pertaining to the project performance aspects (i.e., cost, time, quality, and scope). In this process, the Optimal CRV_i for each aspect has been applied as a reference value. While applying 95% of the confidence, it generates a total of 40,000 samples (10,000 samples from each category, i.e., cost, time,

quality, and scope). Considering data imbalance conditions in the problem at hand, we applied SMOTE to generate a sufficiently large number of data samples. In our proposed work, the SMOTE algorithm generated synthetic samples on the basis of feature space similarities between the existing samples (i.e., risk components and corresponding CRVs) in the minority class. In the proposed model, first, we employed a CRV sample (1) from the dataset and considered its K-nearest neighbor (k-NN) based on Euclidean distance so as to form a vector between the current data points and one of these k-neighbors. The new or updated synthetic data sample was retrieved by multiplying this vector by a random number or weight factor (here, we used four weight factors a, b, c, d to be multiplied with cost, time, quality, and scope, respectively). Notably, the sum of these weight factors was confined as 1. Thus, by multiplying the respective weight factors by each risk element and adding the product to the current data points, we obtained a total of 40,000 samples (2). Thus, the proposed sampling method enabled a balancing of minority class instances and its distribution across the samples (these are those risk components that were given a low rank on the Likert scale but can have a vital impact on project success or performance aspects (i.e., cost, time, quality, and scope).

$$\text{CRV\_Sample} = aA(\text{COST}) + bB(\text{TIME}) + cD(\text{QUALITY}) + dD(\text{SCOPE}) \quad (2)$$

In (2), $\sum(a, b, c, d) = 1$. Additionally, in sync with dynamic socioeconomic and political (contemporary risk changes over project cycle), the weight parameters (a, b, c, d) were varied randomly to generate 40,000 sampled data. Thus, the data prepared can have sufficient information to make learning better and can help in achieving more relevant risk identification toward construction megaprojects.

3.4.3. Feature Selection Based on the Wilcoxon Rank-Sum Test

As discussed in the above section, to retain sufficiently large feature sets (i.e., risk components or factors having an impact on project success or performance), we retrieved a total of 40,000 data elements, constituting features. However, learning over such a significantly large number might force a machine learning method to undergo local minima and convergence. On the other hand, processing a large feature set, irrespective of the level of significance of data elements, can reduce overall accuracy and computing performance. Considering this fact, we performed a significant predictor test, often called the Wilcoxon rank-sum test (WRST), which retains the most significant features having an impact on risk identification. As discussed in previous sections, we used a total of 15 features for each risk factor; hence, the total feature set for 40,000 samples cannot be inevitably significant toward risk characterization or prediction. In other words, not all samples can have a decisive impact on the final prediction output. Moreover, these less significant feature sets often impose substantial computational overhead. Considering this fact, we applied a WRST algorithm to select the most significant features for further computation. The WRST is a type of nonparametric test with independent samples (risk factors). This approach examines the correlation between risk variables and allied samples, as well as their impact probability on the construction project's performance aspects (cost, time, quality, and scope). In our proposed model, the WRST algorithm estimated the correlation between the different features and the respective impact on the aforementioned megaproject's performance variables (i.e., cost, time, quality, and scope). This method obtained a p-value for each feature variable in reference to its significance toward the performance variables. Thus, based on the p-value, each feature element was labeled as significant or insignificant, where the level of significance was assigned as $p = 0.05$. In this manner, only those features and allied instances with a significant correlation were retained to perform further computation. This process primarily focused on retaining significant features to avail a further analytics solution while maintaining low computational complexity.

3.4.4. GA–K-Means-Assisted Risk and Sub-Risk Segmentation

In this research, our key goal is to identify the most decisive risk factor along with sub-risk factors that are highly correlated with key risk factors so that risk distribution and management among the multiple stakeholders can be carried out appropriately. This problem can be easily solved as a clustering problem in which the key risk can be identified as the cluster centroid, while the closest or nearer risk components can be obtained as sub-risk factors. With this motive, once the optimal set of feature elements toward risk identification was obtained, we applied the K-means clustering algorithm to segment the different key risks. However, given that the major classical K-means models often undergo detrimental performance due to random initial centroid assignment, we applied an evolutionary computing algorithm, i.e., a genetic algorithm (GA), for centroid estimation. A detailed discussion of the risk identification based on the proposed GA–K-means clustering algorithm is provided as follows.

K-Means Clustering Algorithm

Typically, K-means clustering is a kind of unsupervised learning approach that groups unlabeled data in multiple clusters or groups. Mainly it focuses on identifying groups within large unstructured data, in which the total number of clusters used to be presented as K. The K-means algorithm operates iteratively to assign each data element (here, the risk components of factors) to one of the K clusters based on the input features provided. In this manner, it clusters the overall data elements into certain groups based on the corresponding feature similarity. K-means clustering estimates the centroid of each cluster, signifying the collection of data elements with similar features or traits. Thus, it generates the centroid for each cluster signifying the definition of each group. In sync with our proposed risk identification system, a centroid states the key risk (decisive or high-risk value component) around which other sub-risks (like connected elements of clusters) can be found. For this reason, in estimating the centroid, signifying high-risk components is vital. On the contrary, the classical K-means algorithm applies random centroid information to perform clustering and can therefore exhibit inaccurate clustering output, especially in our research scenario, where the identification of high-risk value is a must. Realizing this fact, we developed the GA–K-means clustering algorithm, in which, unlike the classical K-means algorithm, a GA was applied as a heuristic model to identify the optimal centroid information. The details of the proposed GA–K-means algorithm are presented in the subsequent section; however, before centroid optimization, understanding the data assignment for initial clustering is a must. Being an iterative refinement-based concept, the K-means algorithm first takes the total number of clusters K and the feature set as input and subsequently exploits data features such as homogeneity heterogeneity and inter-element distance, grouping them into certain distinct clusters. Notably, it clusters the data elements on the basis of inter-element (feature) similarity. The overall clustering method undergoes two sequential steps: first, data assignment, and second, centroid estimation and updates. In a typical data assignment, each element is assigned to its nearest centroid based on distance information such as the squared Euclidean distance, Makowski, and City-Block. While the efficacy of such distance parameters toward optimal clustering remains an open research area, in this research, we applied a different distance estimation method to perform data assignment. The performance with the different distance estimation algorithms is discussed in the next section (Results and Discussion). Considering data assignment using squared Euclidean distance measure, let c_j be the centroids for a feature set C. Each data element x can then be assigned to a cluster based on the condition defined in (3).

$$\arg\min_{c_j \in C} dist\left(c_j, x\right)^2 \tag{3}$$

In (4), *dist* (.) signifies the standard (L_2) (Euclidean) distance; however, in this paper, we applied different methods such as Makowski and City-Block. Considering the set of data points allocated to each *i*th cluster, the centroid estimation is provided as s_i, which

is estimated using a centroid estimation and update method (4). In the centroid update method, the centroid of each cluster is updated dynamically by employing the average of all data elements allocated to that specific (centroid's) cluster. Equation (4) is applied to estimate the centroid of each cluster, iteratively.

$$c_i = \frac{1}{|s_i|} \sum_{x_i \in s_i} x_i \tag{4}$$

The process of the centroid update continues until all data elements are assigned to the most suitable cluster. As discussed above, the classical K-means algorithm applied an average value of the connected elements to decide the centroid but failed to consider inter-element and inter-cluster features to perform cluster enhancement. This could have improved clustering accuracy and the eventual risk segmentation (or clustering) outputs. Considering this as an objective, we applied the GA-based K-means (or GA–K-means) algorithm, in which the GA primarily focuses on exploiting inter-cluster similarity and intra-element associations to perform clustering. In sync with the at-hand research problem, the proposed GA method exploits the relationship between the different risk components pertaining to the different high-risk components and the association between the different high-risk components so as to group different risk factors in different clusters. Before discussing the proposed GA–K-means-based clustering method, a description of the GA algorithm and its significance is provided as follows.

Genetic Algorithm (GA)

A GA is one of the most commonly used evolutionary computing algorithms, derived mainly on the basis of natural events and Darwin's principle of natural selection. Similar to the other heuristic approaches, the GA aims to identify or retrieve an optimal or sub-optimal solution from a large set of solutions. The ability to perform parameter estimation, optimization, and tuning makes the GA one of the best and most used algorithms for solving convex optimization problems. Functionally, the GA applies an objective function, in reference to which it estimates the fitness value of each candidate solution, and reaching a stopping criterion (such as the number of generations), it estimates the eventual solution, often called the best solution, with respect to which a program estimates the final system output or performance. In sync with the at-hand risk identification problem, the GA was used to exploit the different features pertaining to the risk variables and their severity toward a megaproject's performance to cluster them into different groups along with corresponding sub-risk factors.

Functionally, the GA at first initializes a set of chromosomes called the population, signifying a candidate solution. Thus, for each solution, it estimates a fitness value, which is shortened in decreasing order. Again, being a Darwin-principle-based approach, the GA retains only those chromosomes or solutions with a higher fitness value. The low-fitness candidates are subsequently dropped, and hence, it reduces the search space to make computation efficient. To perform controlled reproduction, the GA applied two parameters called crossover and mutation parameters; the former decides what fraction of solutions or candidates will be carried for next-generation crossover, while the latter decides the fraction of candidates or chromosomes to be dropped to retain a productive search space. This process continues with reference to an objective function, also called the cost function, which is reduced over the increasing generation or iterations to yield an optimal solution (post cessation criteria). As stated in this research, the GA is expected to perform a dual task—centroid enhancement and cluster enhancement. Therefore, to enhance the centroid value, it employs a different DB (database) index as an objective function, while for cluster optimization, it uses a silhouette coefficient. The details of the proposed GA–K-means clustering model for risk segmentation or identification are provided as follows.

GA–K-Means Clustering

In the proposed GA–K-means model, clustering can be defined as a stochastic model with data elements known as the population (often called chromosome), where the population $P = \{Ch_1, Ch_2, \ldots, Ch_P\}$. In this method, each chromosome signifies a solution to the clustering problem. Here, the candidate solution Ch_i is estimated by means of its "fitness value," which helps in identifying the best centroid to perform clustering. In the case of an inappropriate centroid, the proposed GA–K-means algorithm performs regeneration of the new population with better-hypothesized fitness values to improve clustering accuracy. The key sequential steps involved in the proposed GA–K-means algorithm implementation are as follows:

I. Binary Data Presentation

In this step, each risk component or sample obtained after WRST feature selection is assigned as a population to estimate high-risk components or the centroid. In the proposed model, we maintained a length of chromosomes equal to the total number of data elements or feature elements, where the *i*th gene of the chromosome represented the *i*th elements in the dataset. Now, for an element i (say, population) to be the centroid for a cluster, the *i*th gene is labeled as "1"; otherwise, it is labeled "0." Here, we selected a value of K in the range $[K_{min}, K_{max}]$, where we assigned K_{min} as 2, while K_{max} was hypothesized to be 15, considering the risk variables and their possible categories. However, in practical cases, the size of K_{max} can be either $l/2$ or $\sqrt{l}$, where l states the chromosome's length.

II. Population Initialization

Consider S_P as the population size, while Ch_{p} is the population encompassing p distinct chromosomes, (i.e., $p = 1, 2, \ldots, S_P$). Here, a non-negative integer value $\mathrm{K_p}$ is randomly chosen from $[K_{min}, K_{max}]$, and thus, the gene associated with the index of the chosen data elements is labeled as "1," while the remaining are labeled as "0." Deploying chromosomes signifying the sub-solution or candidate solution, we estimated the fitness value for each chromosome. In the proposed risk identification model, we focused on improving both the centroid value (i.e., the high-risk component) and the cluster value (i.e., optimally grouped sub-risk values pertaining to an accurately identified high-risk value). Considering this motive, we applied a dual-objective-function concept—one function dedicated for centroid optimization and one for cluster enhancement. More precisely, we applied homogeneity and heterogeneity among the data elements with the different distance functions to perform centroid optimization. On the contrary, we applied the silhouette coefficient as the second objective function to ensure and verify whether the risk components have been assigned to the corresponding closest and dependent high-risk components. The details of these objective functions or the fitness functions are presented as follows.

III. DB-Index Fitness Value Estimation

While the different risk factors in construction megaprojects can have a different level of significance, a few can be highly correlated and have a similar impact on project performance (see Table 1). In this case, assessing features signifying chromosomes (a set of data elements obtained after WRST-based feature selection) for their homogeneity and heterogeneity can help to identify the best chromosome with a higher fitness value. To achieve it, at first, we split input data into small subsets on the basis of homogeneity within the cluster and heterogeneity among the clusters. In this reference, we obtained a DB index, which validates whether the centroid selected has highly connected data elements or vice versa. To achieve it, we estimated the dispersion measure of a cluster C_i, where $i = 1, \ldots, K_r$ is available in the chromosome set Ch_P, using (5).

$$S_{i,q} = \left(\frac{1}{|C_i|} \sum_{x \in C_i} \|x - z_i\|_2^q \right)^{\frac{1}{q}} \quad (5)$$

In (5), $S_{i,q}$ presents the dispersion between the data element i and the feasible centroid candidate q. Here, the centroid of the ith cluster C_i is given as z_i. In the proposed method, the maximum value of intra-cluster similarity was obtained signifying the similarity between the ith cluster C_i and another cluster using (6). Here, the distance vector $d_{ij,t} = d(C_i, C_j)$ applied three different kinds of algorithms: the Euclidean distance, Minkowski, and City-Block algorithms. These algorithms were used distinctly (to assess relative performance) to estimate the distance between clusters C_i and C_j.

$$R_{i,qt} = \max_{j,j \neq i} \left\{ \frac{S_{i,q} + S_{j,q}}{d_{ij,t}} \right\} \tag{6}$$

In (6), $\boldsymbol{t}$ states the order. In this manner, we derived the value of the ***DB*** index DB_P for a chromosome Ch_P by applying (7).

$$DB_P = \frac{1}{k_r} \sum_{i=1}^{k_r} R_{i,qt} \tag{7}$$

With the estimated value of the ***DB*** index for each chromosome of the candidate solution, we obtained the fitness value (8).

$$Fitness(Ch_P) = \frac{1}{DB_P} \tag{8}$$

To enhance the computational efficiency, we assigned q and t as 1 and 2, respectively.

Once the fitness value of each chromosome was calculated, a candidate with the highest fitness value (8) was chosen as the centroid, while other subsolutions with a relatively lower fitness value were considered for reproduction (next generation). Finding a selected chromosome (with the temporarily highest fitness value) with an insufficient fitness value, the GA executes the crossover and mutation process so as to gain a candidate with a better fitness value to cluster risk parameters optimally. In the GA, crossover goals to generate new solutions with a better fitness value are assessed toward its suitability as the centroid. The key goal behind using this crossover method was to achieve the chromosomes or population with better fitness and hence a higher probability of being considered as a centroid value. Thus, the proposed crossover model generates a new chromosome Ch_{new} by manipulating the chromosome with the highest fitness value in a manner such that each centroid candidate is substituted by the data element nearest to the mean center. If Ch_{new} possesses a higher fitness value than Ch_j, then Ch_j is substituted by Ch_{New}. In this manner, a data element (signifying risk factor) with the highest impact or significance toward construction megaproject performance is selected as the centroid. In this manner, it segments the high-risk factors along with the highly correlated sub-risk components, which can be vital for risk management practices.

IV. Silhouette-Coefficient-Based Fitness Estimation

As stated, to serve the dual-objective function, the GA employs DB-index-based centroid optimization, while the silhouette coefficient is applied to improve clustering. In other words, the silhouette coefficient verifies whether each data element (representing risk factors) is clustered in the most appropriate cluster with the optimal centroid. Typically, a higher silhouette coefficient means that the two data points or risk factors are highly correlated or connected. Therefore, estimating this coefficient for each data element can help in aligning the most correlated risk factors to the corresponding (optimal) cluster.

In the GA–K-means model, we estimated the silhouette coefficient by using the average distance between the data elements in the same cluster in comparison to the average distances between data points in other clusters (signifying another high-risk factor). Let K be the cluster encompassing data elements $x(i)$. Similarly, let the average distance between each data element in cluster K be $a_{x(i)}$. Let $b_{x(i)}$ be the minimum average

distance between $x(i)$ and each comprising element in other clusters, which are not the member in cluster K. Thus, the silhouette coefficient of $x(i)$ is obtained using (9).

$$S_{x(i)} = \frac{b_{x(i)} - a_{x(i)}}{max\left(a_{x(i)}, b_{x(i)}\right)} \tag{9}$$

where $x(i)$ represents the data elements available in the cluster, and $i = 1, 2, 3, \ldots n$, $a_{x(i)}$ presents the average distance between each data element and $x(i)$. The other variable $b_{x(i)}$ signifies the minimum average distance between $x(i)$ and each data element in other clusters. With these available attributes, the silhouette average of each cluster can be obtained using (10).

$$S_k = \frac{1}{n}\sum_{i=1}^{n} S_{x(i)} \tag{10}$$

In (10), k presents the number of clusters, and n states the total number of data elements (or sub-risk factors) in the same cluster (with the specific high-risk factor). Now, the silhouette average of all the clusters can be obtained using (11).

$$S_{Avg} = \frac{1}{m}\sum_{k=1}^{m} S_k \tag{11}$$

In (11), m presents the total number of clusters. Thus, when implementing the dual-objective GA–K-means algorithm, all data points signifying risk factors in construction megaprojects were clustered into key high-risk components, followed by highly connected risk subcomponents. Such risk segmentation can be vital for a risk management framework to distribute different risk factors among the different stakeholders. This, as a result, can ensure higher risk avoidance and also minimize the conflict of responsibility in supporting the project's endeavors. Notably, the data considered in this research were primary samples, processed with SMOTE sampling with varying risk impact weightage. It syncs well with dynamic socioeconomic and political conditions, and hence, the eventually segmented high-risk components and allied sub-risks factors can be more vital toward an optimal risk management practice. To validate the performance by the proposed risk assessment and segmentation approach, we obtained performance outputs in terms of the different parameters. The details of the simulation results and allied inferences are given in the subsequent section.

4. Results and Discussion

Considering the significance of earlier risk identification toward construction megaprojects, the key emphasis in this research is on exploiting different risk factors and their impacts on the performance aspects of construction megaprojects. In other words, unlike classical methods, in which either author performs a qualitative study to identify risk variables in midsized projects or megaprojects, we considered a cross, empirical, machine-learning-based approach to identify high-risk components or factors and allied sub-risk components based on their corresponding impact on a project's overall endeavors (cost, time, quality, and scope). Due to the study using a cross, empirical, machine-learning model, we applied an empirical study or a quantitative method to collect expert responses on different key risk factors that have an impact on construction megaproject success, especially in the form of its cost, time, quality, and scope. The respondents were asked to state their opinions toward the impact of 63 different risk factors on a construction megaproject's endeavors or performance aspects. To collect their responses, we applied a five-item Likert scale, where 1 signifies their disagreement of the impact of a specific risk factor (Table 1) on a megaproject's performance, and 5 signifies an undeniable impact on a project's endeavors (cost, time, quality, and scope). In other studies, a quantitative method enabled the identification of different risk factors based on experts' responses; however, the cumulative impact of those risk factors and the interrelation among different risk

factors could not be examined, which could have helped in identifying the most decisive risk factors along with the related sub-risk factors to make risk distribution and allied management more effective. Considering this, we processed the collected primary data in a machine-learning model to predict the risk of higher severity and determine impact significance. Since this problem is similar to a clustering issue (i.e., identifying high-risk components and allied sub-risk factors among the 63 risk variables (Table 1)), we applied K-means clustering as a machine-learning model. However, given the key limitation of the standard K-means clustering algorithm, i.e., the impact of inappropriate cluster centroids on classification accuracy, we applied a GA. A GA can serve a dual purpose: It can improve centroid estimation for K-means and apply the silhouette coefficient as an objective function to ensure that the risk components are optimally clustered with the most relevant high-risk components or cluster centroids. Thus, the use of GA–K-means algorithm was used to achieve accurate high-risk factor identification and most associated sub-risk component segmentation so that those identified risks in different clusters could be assigned across the different stakeholders in megaproject management and execution.

In addition to clustering enhancement, to map the relationship between the different risk variables, their severity, and the corresponding impact on megaproject performance endeavors (in terms of cost, quality, time, and scope), we performed extensive data processing with SMOTE sampling. This subsampling approach obtained a total of 40,000 samples, signifying the different risk variables, with different or distinct severity and inclusive impacts on project performance. In this manner, we employed a sufficiently large amount of data inputs to perform clustering-based risk identification. Thus, implementing our proposed GA–K-means model with quantitatively derived risk-oriented data samples, the proposed model yielded a set of high-risk components and corresponding highly correlated sub-risk factors, which can be applied for real-time risk management practices, especially pertaining to construction megaprojects. To assess performance, we first examined the different distance-based clustering models. To estimate cluster centroids and allied initial connected elements (subcomponents), the K-means algorithm applies different types of distance information, i.e., Euclidean distance, Manhattan, Minkowski distance, and City-Block. Here, we applied the different distance algorithms to cluster the data elements, and the corresponding efficiency was examined in terms of silhouette coefficients. We varied the number of clusters to obtain corresponding silhouette coefficients, in which a higher silhouette coefficient value signified a higher accuracy of the clustering achieved. Results are presented in Tables 3 and 4.

Table 4. Numbers of clusters vs. the silhouette coefficient with the standard K-means clustering algorithm for risk identification.

	Silhouette Coefficient (GA K-Means)			
No. of Clusters	**Euclidian**	**Manhattan**	**City-Block**	**Minkowski**
2	0.618	0.611	0.639	0.618
3	0.626	0.609	0.616	0.611
4	0.689	0.662 *	0.651	0.618
5	0.693 *	0.671	0.669 *	0.660 *
6	0.611	0.613	0.610	0.605
7	0.621	0.600	0.619	0.610
8	0.631	0.644	0.662	0.608
9	0.640	0.683	0.661	0.601
10	0.671	0.628	0.623	0.609

* Highest values.

These results present a clustering efficacy over varying cluster sizes, where, by observing overall performance, one can find that the silhouette coefficient for four clusters, especially with the Euclidean distance measure, yields a better performance (i.e., a higher silhouette coefficient). On the other hand, the above results also revealed that, in comparison to the classical or native K-means-based clustering, the GA–K-means algorithm

exhibited a better performance in terms of a higher silhouette coefficient with the Euclidean distance measure. This confirms the suitability of the proposed GA–K-means clustering algorithm to at-hand risk identification and sub-risk segmentation problems. In sync with the above results (Table 3), for further performance assessment, we applied the Euclidean distance model in (3) to perform further computation.

One of the key objectives of the proposed research was also to identify the key risks pertaining to the "targeted performance aspect" (risk factors pertaining to time, cost, quality, and scope). In other words, considering a typical construction megaproject with a long-term development plan, in which the likelihood of performance dynamism cannot be ignored, we estimated the most critical risk factor with targeted performance aspects, such as preference toward either "timely execution," "cost-efficient development," "quality construction," and "higher success scope over a period." To achieve this, our proposed GA–K-means clustering model identified the different risk variables with distinct targeted performance aspects. Executing the program randomly, we obtained the set of different weight parameters a, b, c and d in (2), with respect to which the proposed model identified the key risk components and allied sub-risk variables. To make the further discussion easier, we redefine the critical risk (2) as (12).

$$= w_{time} \cdot va(k,1) + w_{cost} \cdot va(k,2) + w_{quality} \cdot va(k,3) + w_{scope} \cdot va(k,4) \quad (12)$$

In (12), the different weight parameters including w_{time}, w_{cost}, $w_{quality}$, and w_{scope} presents the project target or endeavors, and with the higher weightage value, the proposed system is expected to provide the set of risk factors to be addressed. In the proposed work, the weight components w_{time}, w_{cost}, $w_{quality}$, and w_{scope} were selected as per project targets or goals, with respect to which the data samples are generated, as discussed in the previous section, and key risk factors including high-risk components and sub-risk components are identified by performing clustering. We simulated our proposed model with the different set of weight values pertaining to w_{time}, w_{cost}, $w_{quality}$, and w_{scope} and obtained the optimal set of risk components to be addressed to achieve that targeted performance aspect. Some of the simulation outputs and their inferences are discussed in the following.

For the above-stated results (Table 5), we assigned w_{time} = 0.0806, w_{cost} = 0.4761, $w_{quality}$ = 0.3269, and w_{scope} = 0.1161. In this case, considering the weightage toward cost-effective construction or performance, a higher weightage was assigned to the cost aspect (w_{cost} = 0.4761), signifying a megaproject with cost efficiency as a project goal. In this case, "delay in obtaining traffic regulation order" was identified as the most critical or high-risk component, while other sub-risk components identified were inappropriate equipment, political and legal issues, political instability, government intervention, and unforeseen circumstances. As indicated in Table 5, the weightage (interest) toward time is at a minimum, signifying that the at-hand project does not focus more on reducing execution time; rather, it targets cost-efficient project completion. In this case, the key risk factors to be considered are segmented, as given in Table 5.

As depicted in Table 6, considering the project goal of time-efficient construction and completion, we simulated our proposed model with different weightage values, where the weight for w_{time} was the highest (0.875) among the other project performance endeavors. The simulation results (Table 6) show that to achieve timely project delivery or completion, retrieving all related public or private acknowledgments and orders, including local governing bodies, traffic, and allied regulation orders, is a must. In major cases, due to the lack of regulatory issues, a megaproject often becomes delayed for a long time. In addition to the regulatory confirmation and regulation orders, the proposed model identifies faulty engineering design, political instability, a lack of political support, inappropriate equipment and materials (Table 6), as risk factors that can prolong a project (especially a megaproject) completion period. However, since a megaproject handling vendor or firm would have sufficient infrastructure with suitable tools and materials, flaws in engineering designs can increase project completion time, as indicated through the proposed model (Table 6).

Table 5. Test Case 1. Cost efficiency.

Weightage	High-Risk Component	Sub-Risk Component
w_{time} = 0.0806 w_{cost} = 0.4761 $w_{quality}$ = 0.3269 w_{scope} = 0.1161	Delay in obtaining temporary traffic regulation orders	Inappropriate equipment and material quality Unforeseen site conditions Incorrect take off calculation Construction and implementation error from faulty design Pollution and vibration Political instability Lack of political support Change in government Government intervention Political and Legal

Table 6. Test Case 2. Time efficiency.

Weightage	High-Risk Component	Sub-Risk Component
w_{time} = 0.875 w_{cost} = 0.6351 $w_{quality}$ = 0.2207 w_{scope} = 0.0567	Delay in obtaining temporary traffic regulation orders	Pollution and vibration Construction and implementation error from faulty design Political instability Lack of political support Change in government Unforeseen site conditions Government intervention Permits and licenses Inappropriate equipment and material quality

From Table 7, in which the weightage toward project scope is the highest among the other performance aspects, it can be found that, in a construction megaproject, inadequate preliminary surveys and site information are high-risk factors. Undeniably, a megaproject involving billions of investments requires optimal surveys regarding the locality, regional demands, the ease of construction and site reachability, and future endeavors. Therefore, inappropriate assessments and allied decisions might force a megaproject to undergo significant loss. On the contrary, to enhance the scope of a megaproject, retaining better reachability or connectivity, meeting local or regional demands, acquiring suitable quality-oriented material and equipment can help in achieving target performance. In sync with the targeted scope as a performance aspect, the proposed risk identification model showed that optimal risk management policies or practices require addressing unforeseen project change problems, changes in government or local bodies, government interventions, pollution and vibration, implementation (here, construction) errors. Thus, the risk management team must address these risk factors (Table 7) if it aims to retain a better project scope.

Table 8 presents the risk identified when targeting quality-centric construction megaproject completion. When giving the highest priority or preference to the quality aspect, the proposed prediction model identified "Inappropriate equipment and material quality" as the key risk factor, demanding its enhancement in achieving target endeavors. In addition to equipment and material quality as a risk factor, the proposed model identifies other similar quality constructs that need to be addressed. These sub-risk factors are machinery failure risk, unforeseen site conditions, poor equipment performance, unorganized soil or local ground conditions, implementation errors, design errors, and multiple modifications. Thus, a risk management team of a construction megaproject needs to address these key risk factors to accomplish a quality construction outcome. Thus, our proposed GA–K-means clustering model identified the different kinds of high-risk factors and closely related sub-risk components, the addressing of which can lead to better project outcomes in terms of timely project completion, quality construction, cost-efficient construction, and construction with a better scope. Addressing the allied risk factors across the project cycle can help in achieving targeted performance endeavors (cost, time, quality, and scope).

Table 7. Test Case 3. Project scope.

Weightage	High-Risk Component	Sub-Risk Component
w_{time} = 0.3299 w_{cost} = 0.2694 $w_{quality}$ = 0.0335 w_{scope} = 0. 3671	Inadequate preliminary survey and site information	Unforeseen modification to project scope
		Construction and implementation error from faulty design Pollution and vibration Inadequate preliminary survey and site information Political instability Inadequate environmental analysis Demands of local people Lack of political support Change in government
		Unforeseen site conditions Government intervention Inappropriate equipment and material quality

Table 8. Test Case 4. Project quality.

Weightage	High-Risk Component	Sub-Risk Component
w_{time} = 0.2349 w_{cost} = 0.2162 $w_{quality}$ = 0.3446 w_{scope} = 0.2044	Inappropriate equipment and material quality	Machinery Failure/breakdown Unforeseen site conditions Poor Equipment performance Low skilled/incompetent workforce Poor site coordination/work organization Unrecognized soil structure/unforeseen ground condition Construction and implementation error from faulty design Incompetency of Designers Inadequate design and design errors
		Unforeseen multiple modifications to project scope

Research Question Reasoning

Regarding RQ1 and the research outcomes (Tables 4–8), a megaproject is less affected due to the impulsive risk factors or the short-term risk factors. On the contrary, it often is influenced by long-term decisions inculcating over the complete project cycle. A small or midsize project, which often has limited investment and time, is relatively less affected by government changes, policy changes, local regional support and changes, cost increases. On the contrary, a megaproject, especially in construction, continues for a long period, even for decades or at least 4–5 years. Undeniably, over such a long time, the cost of materials, equipment, resources, policies, the government is often changed. These factors, as a cumulative risk component, influence the performance of a construction megaproject. On the contrary, small or midsized construction projects are less affected. These facts affirm the acceptability of RQ1. Considering RQ2, which is intended to assess the key risk factors that often affect a project's endeavors, such as its cost, time of completion, quality, and scope, the results obtained in Tables 3, 4, 5, 6 and 7show that legal or regulatory support on time (time), government changes and construction errors (cost), inferior materials and equipment, inappropriate designs (quality), and redesigns with inferior construction quality (scope) are the key high-risk factors in construction megaprojects. The overall results and their respective impacts on a construction megaproject's risk management affirm that the use of expert inputs and machine-learning-based high-risk identification, followed by sub-risk identification, is an appropriate tool for risk management in construction megaprojects. This affirms the acceptability of RQ3. These results affirm the acceptability of RQ4; moreover, signifying the use of descriptive analysis of expert consensus on the risk severity feature with the Wilcoxon rank-sum test as a feature selection method and the multiconstraint (cost, time, quality, and scope), objective-function-based GA–K-means algorithm yields optimal risk identification and segmentation in construction megaprojects. The comparison of the performance outcomes in Tables 2 and 3 affirm that, unlike classical

K-means clustering, the use of heuristic driven GA–K-means clustering can yield more reliable results (due to higher silhouette coefficients with four clusters, signifying four project goals—time, cost, quality, and scope) and is a more effective risk identification solution for construction megaprojects. This also confirms the acceptance of RQ5.

5. Conclusions

In this paper, key emphasis was placed on exploiting expert-response-driven data analysis to predict high-risk factors and the corresponding closely related sub-risk components pertaining to construction megaprojects.

Unlike classical methods, in which authors have either applied qualitative or quantitative approaches to identify key risk factors in projects, to avoid any possibility of biasing impact, this research exploited the efficiency of different state-of-the-art statistical analysis tools and machine learning and artificial intelligence concepts. It performed an analysis of experts' perceptions toward the different risk factors having an impact on project endeavors or aspects such as time, cost, quality, and scope. It also exploited experts' perceptions as input data to mine the associations among the different risk components and clustered them together to determine high-risk factors with decisive impacts on project endeavors (i.e., time, cost, quality, and scope), along with the highly correlated risk factors. Thus, this identification of high-risk factors and related sub-risk factors can enable risk management teams or project management teams to distribute risk optimally across different stakeholders to make optimal risk avoidance measures.

To alleviate the impact of any possible biasing (of the experts' responses), the proposed work performed a descriptive analysis of the samples collected, which was followed by data subsampling. The proposed model performs SMOTE sampling, which distributes and generates a sufficiently large number of samples with a corresponding impact on project endeavors. The use of weightage adaptive sample generation or subsampling enabled the retrieval of samples over varying dynamic project conditions and expectations. Thus, with the prepared large risk factors (in sync with the different project aspect), this research applied GA-based K-means clustering, i.e., GA–K-means, with the Euclidean distance method and dual-objective functions of the silhouette coefficient (for cluster enhancement and verification) and the DB index (for centroid optimization) to identify high-risk factors, along with highly correlated or closely related sub-risk factors.

Considering the eventual outcome of the proposed model, for a construction megaproject to achieve cost-efficient performance, as shown in Table 5, delay in obtaining traffic regulation order is identified as a high-risk factor. The related sub-risk components are inappropriate equipment, political and legal issues, political instability, government intervention, and unforeseen circumstances, which are expected to be addressed optimally by the allied risk management team to accomplish overall project endeavors. Similarly, factors such as regulatory confirmation and regulation order delays, wrong engineering designs, political instability, a lack of political support, and inappropriate equipment and materials turned out to be key risk factors influencing the time efficiency of a construction megaproject.

A construction megaproject, being a far-long targeted infrastructure, demands the addressing of different risk adversaries to retain a better scope. This study revealed that inadequate preliminary surveys and site information are high-risk factors impacting the scope of a megaproject. Retaining better reachability or connectivity, meeting local or regional demands, and acquiring suitable quality-oriented material and equipment are needed to achieve target performance.

The proposed risk identification model also showed that optimal risk management policies or practices require the addressing of unforeseen project-change problems, changes in government or local bodies, government interventions, pollution and vibration, and implementation (here, construction) errors. The research also showed that equipment and material quality is a high-risk factor affecting the quality of construction, (as shown in Table 8) and machinery failure risks, unforeseen site conditions, poor equipment performance, unorganized soil or local ground conditions, implementation errors, design errors, and

multiple modifications also impact quality. Therefore, a risk management team must handle these key risk factors to achieve target performance with cost efficiency, timely project accomplishment, quality construction, and better project scope.

However, as with every other research using human inputs in the form of expert opinion and interviews, this research had the limitation of efficiently obtaining inputs from experts since it was often difficult to convince the participants regarding the importance of their inputs. The focus group consensus was also difficult to achieve, and sometimes, there were instances of missed appointments and rescheduling due to the busy schedule of the focus group participants.

This contribution (risk identification in construction megaproject) could be vital for construction firms, allied decision makers, and risk assessment and allied strategic solution providers to make optimal dynamic or proactive decisions to achieve various project endeavors. The study could also be extended for future endeavors in applying similar models for other forms of the construction industry such as public–private partnership (PPPs) and on medium- and small-scale projects.

Author Contributions: The present research was conceptualized and designed by J.P. The data collection, data analysis, manuscript writing was performed by D.B.C. Data interpretation, Editing and interpretation of results were conducted by R.M.R.P. The overall supervision and result interpretation and Analysis were conducted by J.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding and was solely prepared by the authors.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data is available from the corresponding author on request.

Acknowledgments: We thank the Dean of the School of Civil Engineering, Vellore Institute of Technology, for supporting our research.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A Questionnaire

Appendix A.1 Sociodemographic Variables

Q1. Please specify your gender.

- Male
- Female

Q2. Please specify your age range.

- 21 years to 30 years
- 31 years to 40 years
- 41 years to 50 years
- >50 years

Q3. Please specify your role/designation in the project.

- Project Manager
- Contractor/Vendor
- Consultant
- Supplier
- Others

Q4. Please specify your total years of experience in construction industry.

- 0–5 years
- 6–10 years
- 11–15 years

- 16–25 years
- >25 years

Q5. Please specify the project phase for your involvement.

- Initiation and Planning
- Execution
- Monitoring and Control
- Closure
- Others

Q6. Please specify the project size.

- ≤Rs. 100,00,000
- Rs. 100,00,001–Rs. 500,00,000
- Rs. 500,00,001–Rs. 1000,000,000
- >Rs. 100,000,000

Q7. Is there any risk associated with megaproject delivery that alters the project performance in terms of time, cost, quality, and scope?

- Yes
- No

Q8. Based on your experience, what do you think about the following risk categories impacting the project performance in terms of time, cost, quality, and scope for a megaproject?

	Risk Categories	Strongly Agree (5)	Agree (4)	Neutral (3)	Disagree (2)	Strongly Disagree (1)
1	Execution Risks					
2	Construction Risks					
3	Technical, Engineering, and Design Risks					
4	Economic and Financial Risks					
5	Social Risks					
6	Environmental Risks					
7	Political and Legal Risks					

Appendix A.2 Project Performance: Risk Impact Identification on Project Objectives

Listed below are the risk factors identified under various risk categories for megaprojects. Please select your response based on your experience to rate the impact of these risk factors on project objectives of time, cost, quality, and scope.

	Risk Factors	Rate (1 to 5) Impact of Risk Factors on			
		Time	Cost	Quality	Scope
	Execution				
1	Utility diversion				
2	Inappropriate equipment and material quality				
3	Permits and licenses				
4	Poor equipment performance				
5	Machinery failure/breakdown				
6	Unforeseen site conditions				
7	Incorrect take off calculation				
8	Delayed supply of material and equipment				
9	Delay in obtaining working drawings/reports/designs				
10	Low skilled/incompetent workforce				
11	Unavailability of materials, equipment, and labor				
12	Delay in obtaining temporary traffic regulation orders				

	Risk Factors	Rate (1 to 5) Impact of Risk Factors on			
		Time	**Cost**	**Quality**	**Scope**
	Construction				
13	Poor site coordination/work organization				
14	Construction failure				
15	Land acquisition for ROW				
16	Inadequate preliminary survey and site information				
17	Unrecognized soil structure/unforeseen ground condition				
18	Delay in transport of ready-mix concrete (RMC)				
19	Construction and implementation error from faulty design				
20	Changes in material during construction				
21	Deviations between specification and implementation				
22	Supply chain breakdown/improper equipment and material quality				
23	Site inaccessibility				
24	Lack of site security for personnel and asset				
	Technical, Engineering and Design				
25	Incompetency of designers				
26	Design changes				
27	Inadequate design and design errors				
28	Modification to drawing/design				
29	Unforeseen multiple modifications to project scope				
30	Delay in obtaining preliminary drawings/reports				
31	Revision in design standard				
32	Inadequate project complexity analysis				
	Economic				
33	Inflation				
34	Foreign exchange rate and interest rate fluctuation				
35	Changes in market conditions				
36	Changes in taxes				
37	Incorrect cost estimate				
38	Financial difficulties/failure of subcontractor				
	Environmental				
39	Natural Disaster				
40	Adverse weather condition				
41	Pollution and vibration				
42	Geology, soil, and topography				
43	Drainage pattern				
44	Inadequate environmental analysis				
45	Land cover (grass, asphalt, trees, water bodies)				
46	Presence of quarries and mines				
	Social				
47	Demands of local people				
48	Public objections				
49	Social issues (tree cutting, shrine removal)				
50	Cultural and heritage sights				
51	New stakeholders with change request				
52	Damage to property and persons				
53	Multilevel decision-making bodies				
	Political and Legal				
54	Changing government regulations/funding policy				
55	Lack of moderators				
56	Legal disputes				
57	Political instability				
58	Changes in local laws and standards (tax imposition)				
59	Lack of political support				
60	Political indecision				
61	Change in government				
62	Multilevel decision-making by government bodies for consent and approvals				
63	Government intervention				

References

1. Miller, R.; Lessard, D.R. *The Strategic Management of Large Engineering Projects Shaping Institutions, Risks and Governance*; MIT Press: Cambridge, MA, USA, 2001; ISBN 0262526980.
2. Zhai, L.; Xin, Y.; Cheng, C. Understanding the value of Project Management from a Stakeholder's Perspective: Case Study of Mega-Project Management. *Proj. Manag. J.* **2009**, *40*, 99–109. [CrossRef]
3. Eweje, J.; Turner, R.; Muller, R. Maximizing Strategic Value from megaprojects: The influence of information-feed on decision-making by the project manager. *Int. J. Proj. Manag.* **2012**, *30*, 639–651. [CrossRef]
4. Floricel, S.; Miller, R. Strategizing for Anticipated Risks and turbulence in large-scale engineering projects. *Int. J. Proj. Manag.* **2001**, *19*, 445–455. [CrossRef]
5. Kardes, I.; Ozturk, A.; Cavusgil, S.; Cavusgil, E. Managing global megaprojects: Complexity and risk management. *Int. Bus. Rev.* **2013**, *22*, 905–917. [CrossRef]
6. Flyvbjerg, B.; Bruzelius, N.; Rothengatter, W. *Megaprojects and Risk: An Anatomy of Ambition*, 1st ed.; Cambridge University Press: Cambridge, UK, 2003. [CrossRef]
7. Vidal, L.A.; Marle, F. Understanding project complexity: Implications on project management. *Kybernetes* **2008**, *37*, 1094–1110. [CrossRef]
8. Koppenjan, J. The formation of public-private partnerships: Lessons from nine transport infrastructure projects in the Netherlands. *Public Adm.* **2005**, *83*, 135–157. [CrossRef]
9. Dimitriou, H.T.; Ward, E.J.; Wright, P.G. Mega transport projects—Beyond the "Iron Triangle": Findings from the OMEGA research programme. *Prog. Plan.* **2013**, *86*, 1–43. [CrossRef]
10. Clegg, S.; Pitsis, T.; Rura-Polley, T.; Marosszeky, M. Governmentality matters: Designing an alliance culture of inter-organizational collaboration for managing projects. *Organ. Stud.* **2002**, *23*, 317–337. [CrossRef]
11. Esty, B. Why study large projects? An introduction to research on project finance. *Eur. Financ. Manag.* **2004**, *10*, 213–224. [CrossRef]
12. Palma, A.; Picard, N.; Andrieu, L. Risk in transport investments. *Netw. Spat. Econ.* **2012**, *12*, 187–204. [CrossRef]
13. Dey, P.K. Project risk management using multiple criteria decision-making technique and decision tree analysis: A case study of Indian oil refinery. *Prod. Plan. Control.* **2012**, *23*, 903–921. [CrossRef]
14. Dey, P.K. Managing project risk using combined analytic hierarchy process and risk map. *Appl. Soft Comput.* **2010**, *10*, 990–1000. [CrossRef]
15. Project Management Institute (PMI). *Organizational Project Management Maturity Model, OPM3®*, 2nd ed.; Project Management Institute: Newtown Square, PA, USA, 2008.
16. Project Management Institute (PMI). *A Guide to the Project Management Body of Knowledge (PMBOK® Guide)*, 5th ed.; Project Management Institute: Newtown Square, PA, USA, 2013.
17. Project Management Institute (PMI). *The Standard for Program Management*, 3rd ed.; Project Management Institute: Newtown Square, PA, USA, 2013.
18. Greiman, V.A. *Megaproject Management: Lessons on Risk and Project Management from the Big Dig*; Wiley: Hoboken, NJ, USA, 2013; ISBN 978-1-118-41634-1.
19. Lehtiranta, L. Risk perceptions and approaches in multi-organizations: A research review 2000–2012. *Int. J. Proj. Manag.* **2014**, *32*, 640–653. [CrossRef]
20. Wu, D.; Olson, D.L. Supply chain risk, simulation, and vendor selection. *Int. J. Prod. Econ.* **2008**, *114*, 646–655. [CrossRef]
21. Dunovic, I.B.; Radujkovic, M.; Vukomanovic, M. Risk register development and implementation for construction projects. *Gradevinar* **2013**, *65*, 23–35. [CrossRef]
22. Marcelino-Sadaba, S.; Perez-Ezcurdia, A.; Lazcano, A.M.E.; Villanueva, P. Project risk management methodology for small firms. *Int. J. Proj. Manag.* **2014**, *32*, 327–340. [CrossRef]
23. Fiori, C.; Kovaka, M. Defining megaprojects: Learning from construction at the edge of experience. In *Construction Research Congress 2005: Broadening Perspectives*; American Society of Civil Engineers: Reston, VA, USA, 2005.
24. Boateng, P. A Dynamic Systems Approach to Risk Assessment in Mega Projects. Ph.D. Thesis, Heriot-Watt University, Edinburgh, UK, September 2014.
25. Bhandari, M.; Gayakwad, P.G. Management of Risk in Construction Projects in Maharashtra. *Int. J. Eng. Sci. Invent.* **2014**, *3*, 14–17.
26. Debalina, B.C.; Jagadeesh, P. Mega-Project Risk Prediction in Construction Management. *J. Crit. Rev.* **2020**, *7*, 2674–2687. [CrossRef]
27. Flyvbjerg, B.; Skamris, H.M.K.; Buhl, S.L. What causes cost overrun in transport infrastructure projects? *Transp. Rev.* **2004**, *24*, 3–18. [CrossRef]
28. Kwak, Y.H. Perceptions and practices of project risk management: Aggregating 300 project manager years. In Proceedings of the Project Management Institute Global Congress North America, Baltimore, MD, USA, 18–25 September 2003; pp. 18–25.
29. Boateng, P.; Chen, Z.; Ogunlana, S.; Ikediashi, D. A system dynamic approach to risks description in mega-projects development. *Organ. Technol. Manag. Constr. Int. J.* **2012**, *4*, 593–603. [CrossRef]
30. Irimia-Dieguez, A.; Sanchez-Cazorla, A.; Alfalla-Luque, R. Risk Management in Mega-projects. *Procedia Soc. Behav. Sci.* **2014**, *119*, 407–416. [CrossRef]

31. He, Q.; Luo, L.; Chan, A.P. Measuring the complexity of mega construction projects in china-A fuzzy analytic network process analysis. *Int. J. Proj. Manag.* **2015**, *33*, 549–563. [CrossRef]
32. Kim, S.G. Risk performance indexes and measurement systems for mega construction projects in China-A fuzzy analytic network process analysis. *J. Civil Eng. Manag.* **2010**, *16*, 586–594. [CrossRef]
33. Alsharef, A.F.A. *Design of a Construction Expenditure Forecasting and Monitoring Tool for NCDOT Megaprojects*; North Carolina State University: Raleigh, NC, USA, 2015.
34. Young, T.L. *Successful Project Management*, 3rd ed.; Kogan Page Limited: London, UK, 2010; ISBN 978-0-7494-5917-8.
35. Cooper, R.; Edgett, S.; Kleinschmidt, E. Portfolio management for new product development: Results of an industry practices study. *R D Manag.* **2001**, *311*, 361–380. [CrossRef]
36. Cooper, D.; Grey, S.; Raymond, G.; Walker, P. *Project Risk Management Guidelines: Managing Risk in Large Projects and Complex Procurements*; Wiley: Hoboken, NJ, USA, 2005; ISBN 978-0-470-02282-5.
37. Torok, R.; Nordman, C.; Lin, S. *Clearing the Clouds: Shining a Light on Successful Enterprise Risk Management*; Executive Report; IBM Global Business Services: Armonk, NY, USA, 2011.
38. Thamhain, H.J. Leading technology-intensive project teams. In Proceedings of the PMI Global Congress 2003-North America, Baltimore, MD, USA, 20–23 September 2003; Project Management Institute: Newtown Square, PA, USA, 2003.
39. Thamhain, H.J. Managing risks in complex projects. *Proj. Manag. J.* **2013**, *44*, 20–35. [CrossRef]
40. Thamhain, H.J. *Managing Technology-Based Projects: Tools, Techniques, People and Business Processes*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2014. [CrossRef]
41. Thamhain, H.J.; Wilemon, D. Building effective teams in complex project environments. *Technol. Manag.* **1998**, *5*, 203–212.
42. Altshuler, A.A.; Luberoff, D.E. *Mega-Projects: The Changing Politics of Urban Public Investment*; Brookings Institution Press: Washington, DC, USA, 2003; ISBN 13-978-0815701293.
43. Mojtahedi, S.M.H.; Mousavi, S.M.; Aminian, A. Fuzzy group decision making: A case using FTOPSIS in mega project risk identification and analysis concurrently. In Proceedings of the 2008 IEEE International Conference on Industrial Engineering and Engineering Management, Singapore, 8–11 December 2008; pp. 1–5.
44. Atkinson, R.; Crawford, L.; Ward, S. Fundamental uncertainties in projects and the scope of project management. *Int. J. Proj. Manag.* **2006**, *24*, 687–698. [CrossRef]
45. Harvett, C.M. A Study of Uncertainty and Risk Management Practice Related to Perceived Project Complexity. Ph.D. Thesis, Bond University, Queensland, Australia, 2013.
46. Toor, S.U.R.; Ogunlana, S.O. Beyond the 'iron triangle': Stakeholder perception of key performance indicators (KPIs) for large-scale public sector development projects. *Int. J. Proj. Manag.* **2009**, *28*, 228–236. [CrossRef]
47. Ke, Y.; Wang, S.; Chan, A.P.C.; Lam, P.T.I. Preferred risk allocation in China's public-private partnership (PPP) projects. *Int. J. Proj. Manag.* **2010**, *28*, 482–492. [CrossRef]
48. Nelson, R.R. IT project management: Infamous failure, classic mistakes, and best practices. *Mis Q. Exec.* **2007**, *6*, 67–78.
49. Van Marrewijk, A.H.; Clegg, S.R.; Pitsis, T.; Veenswijk, M.B. Managing public private megaprojects: Paradoxes, complexity, and project design. *Int. J. Proj. Manag.* **2008**, *26*, 591–600. [CrossRef]
50. Rezakhani, P. A review of fuzzy risk assessment models for construction projects. *Slovak J. Civ. Eng.* **2012**, *20*, 35–40. [CrossRef]
51. Taroun, A. Towards a better modelling and assessment of construction risk: Insights from a literature review. *Int. J. Proj. Manag.* **2014**, *32*, 101–115. [CrossRef]
52. Bruzelius, N.; Flyvbjerg, B.; Rothengatter, W. Big decisions, big risks: Improving accountability in mega projects. *Transp. Policy* **2002**, *9*, 143–154. [CrossRef]
53. Little, R.G. The emerging role of public private partnerships in mega-project delivery. *Public Work. Manag. Policy.* **2011**, *16*, 240–249. [CrossRef]
54. Bing, L.; Akintoye, A.; Edwards, P.J.; Hardcastle, C. The allocation of risk in PPP/PFI construction projects in the UK. *Int. J. Proj. Manag.* **2005**, *23*, 25–35. [CrossRef]
55. Rolstadas, A.; Johansen, A. From protective to offensive project management. In Proceedings of the PMI Global Congress 2008—EMEA, St. Julian's, Malta, 19 May 2008.
56. Krane, H.P.; Olsson, N.O.E.; Rolstadas, A. How project manager-project owner interaction can work within and influence project risk management. *Proj. Manag. J.* **2012**, *43*, 54–67. [CrossRef]
57. Westney, R.E.; Dodson, K. CAPEX VaR: Key to improving predictability. *World Energy* **2006**, *9*, 134–138.
58. Krane, H.P.; Rolstadas, A.; Olsson, N.O.E. Categorizing Risks in Seven Large Projects-Which Risks Do the Projects Focus On? *Proj. Manag. J.* **2010**, *41*, 81–86. [CrossRef]
59. Turner, J.R.; Muller, R. The Project Manager's Leadership Style as a Success Factor on Projects: A Literature Review. *Proj. Manag. J.* **2005**, *36*, 49–61. [CrossRef]
60. Wilkinson, D.; Birmingham, P. *Using Research Instruments: A Guide for Researchers*; Routledge Falmer: London, UK, 2003.
61. Hsieh, H.F.; Shannon, S.E. Three Approaches to Qualitative Content Analysis. *Qual. Health Res.* **2005**, *15*, 1277–1288. [CrossRef] [PubMed]

62. Andric, J.M.; Wang, J.; Zou, P.X.W.; Zhang, J. Fuzzy Logic based Method for Risk Assessment of Belt and Road Infrastructure Projects. *J. Constr. Eng. Manag.* **2019**, 145. [CrossRef]
63. Debalina, B.C.; Jagadeesh, P. An Application of Critical Risk Identification Based on Fuzzy Inputs for Indian Construction Megaproject. *Int. J. Adv. Sci. Technol.* **2020**, *29*, 15209–15234. Available online: http://sersc.org/journals/index.php/IJAST/article/view/32491 (accessed on 15 January 2021).
64. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]