Politecnico di Torino

BOSTON UNIVERSITY

Politecnico di Torino

Corso di Laurea Magistrale in Ingegneria Gestionale

A./A. 2022/2023

Sessione di Laurea Dicembre 2023

Innovative Pre-Tender Costs Estimating Techniques in Construction Projects

Relatore:

Prof. Andrea De Marco

Co-Relatori:

Prof. Filippo Maria Ottaviani

Prof. Jason Wong

Candidata:

Beatrice Audisio

Abstract

The purpose of this thesis is to present a comprehensive overview of innovative pre-charter estimating techniques along with the costs and factors impinging on these estimates for construction projects. Specifically, a foundational introduction to cost estimates and the emerging influence of AI in project management practices will be presented along with the methodology, the PRISMA approach, adopted for this thesis. Following this introductory section, the thesis will present the cost categories of construction projects along with an analysis of their major factors. After this preliminary section, new estimating techniques, categorized into Machine Learning, Knowledge-Based Systems, and Evolutionary Systems will be introduced and deeply analyzed. The presented AI-based estimating algorithms display increased accuracy compared to traditional estimates, justified by their adaptability and ability to detect correlations between variables, thus enabling the representation of complex systems. Effectively, for every technique presented, a general description followed by an example of its implementation from a research paper will be analyzed. The new techniques showcased are all applied through algorithms, specifically, Artificial Neural Networks, Case-Based Reasoning, Genetic Algorithms, and Hybrid models will be presented, as literature review revealed their pivotal role in research. To conclude, this thesis's overarching aim is to provide a clear guide to these new AI-based methods of cost estimation for construction projects, given the escalating popularity of AI practices and the complexity and significant loss of investment in the construction industry due to misestimates.

Keywords: cost estimates in construction projects, AI in project management, estimating techniques, machine learning, artificial neural networks, knowledge-based systems, case-based reasoning, evolutionary algorithms, genetic algorithms.

Table of Contents

Introduction and Methodology

Introduction

The global construction industry, worth at present 14.4 trillion USD (*Global Construction Market Report And Strategies To 2032*, 2023), accounts for 14% of the global GDP and is characterized by a significant number of projects sporting delays and cost overruns, along with conflicts amongst stakeholders. Indeed, according to an interesting review by Atapattu et al., a mean cost overrun of 28% afflicts construction projects (Atapattu et al., 2023), especially in the initial and design stages. Specifically, the main causes of this phenomenon found are: misestimates, inaccurate risk identification, inaccurate scoping, frequent changes, delays, design errors, inaccurate monitoring, scheduling errors, reworks, unsuited construction methodology, overall poor management skills, external factors (Atapattu et al., 2023). The high prevalence of these issues can be attributed to three main factors: the unique and very uncertain nature of construction projects; the fragmented and highly competitive nature of the construction industry; and the ever-increasing challenges facing the industry. An interesting review from McKinsey reports that, following the COVID-19 pandemic, which acted as an accelerating factor, the industry has been disrupted by "A combination of sustainability requirements, cost pressure, skills scarcity, new materials, industrial approaches, digitalization, and a new breed of player" (Ribeirinho et al., 2020). Due to these escalating disruptions, efficient management becomes a key to the success of any construction organization, and as such, it is paramount to recognize the new computation tools that are now becoming more available. Specifically, given the significant financial loss derived by misestimates, the author aims to question the existing state of pre-tender estimation costs in relation to the existing computational skills now available. Additionally, the literature review conducted is employed to formulate a clear and coherent guide to the main AI algorithms that can be now employed for cost estimates in construction projects.

This review will cite other existing research papers with practical applications on historical data, comparisons of the accuracy of different algorithms, and reviews of the state of the diffusion of AI-based techniques in project management and cost estimation. Moreover, general information regarding the functioning of the algorithms and techniques discussed will be presented. For a more well-rounded paper, a chapter illustrating cost estimation's categories and their influential features is included. While this may seem a redundant review, the author aims to provide a distinct compendium of the estimating problem and the AI-based techniques, in order to help managers understand algorithms and their strengths and weaknesses, so as to choose the most suitable one for each instance. With this overall aim in mind, the structure of this paper may not completely resemble traditional systematic reviews, as the main objective was to provide an overview and illustration of the algorithms and procedures employed, along with presenting their accuracy.

The literature research has been conducted adopting the PICOC (Population, Intervention, Comparator, Outcome, and Context) method (Butler, n.d.). As population, research papers focused on the study of cost estimation for construction projects have been chosen. The application of AI-based techniques is the intervention factor, while the comparator is the accuracy of traditional estimates. The outcome is to prove the superior accuracy of AI techniques in the context of construction projects' pre-tender cost estimates. The context has been defined as academic papers in the field of construction projects, from residential buildings to megaprojects such as railroad systems.

According to Turner and Cochrane, construction projects are traditionally classified as Type 1 or Engineering projects, characterized by goals and methods clearly defined since their launch (Turner & Cochrane, 1993). These projects are heavily resource-intensive and activity-based. Hence, whereas they are more likely to succeed than other research or product development-type projects, they bring a higher inherent risk level due to their complexity, uniqueness, uncertainty, major capital outlays, and operating environments (Turner & Cochrane, 1993). This is the rationale behind the choice of construction projects for the investigation of AI-based algorithms for cost estimation. The

author believes that given the significant amount of loss of investment due to misestimates and the traditional underperformance of the construction sector (Ribeirinho et al., 2020), this industry would particularly benefit from the application of these methods.

A first review revealed that the literature is rich in papers analyzing the accuracy of AI-powered estimation techniques with historical data, but there is little evidence of the widespread implementation of AI-based techniques to estimate pre-tender costs during active projects. While there is proof of PM software using machine learning and big data, there is little to no evidence of implementation of AI-powered cost estimates in construction projects, even if there is confirmation of AI-based tools used especially in the testing and simulation phase for megaprojects, like the recent Elizabeth Line in London (Scholl-Sternberg, n.d.). Unfortunately, this is mainly due to the fact that only in the last few years AI services' popularity started to exponentially increase (Artificial Intelligence Market Size/Revenue Comparisons 2022, n.d.), and the fact that at the present moment, while some higher institutions acknowledge the disruptiveness of AI for project management (Ludden, 2019), the majority of universities still focus their cost estimation classes on traditional techniques.

The Role of AI in Project Management

Every year, the investment in new projects is worth 48 trillion dollars. Nevertheless, the success rate of these investments is only 35% (Hastie & Wojewoda, 2015). According to a recent article in the Harvard Business Review, "One reason […] why project success rates are so poor is the low level of maturity of technologies available for managing them. Most organizations and project leaders are still using spreadsheets, slides, and other applications that haven't evolved much over the past few decades" (Nieto-Rodriguez & Viana Vargas, 2023). These methods are still useful to track the development of a project and measure its deliverables and deadlines, but they are not adaptive tools, thus clashing with the overly adaptive and changing world we are living in nowadays.

By 2030, Gartner has foreseen that up to 80% of the project management tasks will be taken over by AI through big data, natural language processing, and machine learning (Costello, 2019). Indeed, AI-based software is quietly disrupting the project management world and several areas will be impacted upon, leading to a radical shift in the role of project manager.

Nowadays, AI software is already automating operations such as scheduling activities and follow-ups, the creation and distribution of reports, verifying policies' compliance, and gathering feedback. These systems integrate with popular tools like JIRA and Slack, increasing project managers' efficiency with their ability to process and manage patterns combined with their superior computational prowess. Project managers spend up to 90% of their time communicating with a project's stakeholders. This means ensuring effective and continuous communication throughout the project lifecycle. Indeed, "communications remain one of the major differentiators between project success and failure" (Kliem, 2007). Emails and follow-ups, along with check-ins are a significant part of the communication processes in a project. These are routine procedures that can be automated and handled by an AI, freeing project managers from these repetitive tasks. According to the McKinsey Global Institute analysis, on average, 28% of work time is spent on email (Plummer, 2019). This is an important timesaver, which enables project managers to focus on more complex tasks behind projects and their employees, dedicating more time to in-person communication that according to a recent study by Florida International University College of Business is more productive than online communication. Indeed, "Negotiating or working together to solve a problem is more difficult over email or instant messenger than working in person because text-based communication limits visual, vocal, and nonverbal cues"(Corzo, 2022).

Outside the realm of communication, AI is slowly being introduced to other areas of project management. An important one is project risk evaluation and management. New software help PM anticipate risks with big data and machine learning, proposing mitigation actions, and soon, as they

are adaptive systems, they will be able to automatically address common risks automating a significant part of risk management processes (Nieto-Rodriguez & Viana Vargas, 2023).

AI will lead to improved project scoping by automating the time-consuming collection and analysis of user stories, bringing increased accuracy, and eliminating ambiguities and omissions (Nieto-Rodriguez & Viana Vargas, 2023). Moreover, it will automate scheduling processes, reporting, and even the draft of concept notes. In fact, ChatGPT revealed the incredible asset that can be AI when analyzing databases; and even now it can already generate concept notes (OpenAI, 2023). AI is also at present in use through virtual assistants, like PMOtto, that through machine learning can assist users in converting their natural language into actions within their project management software. Being an adaptive system, it already expresses recommendations based on similar projects and available information (*PMOtto.Ai – Ricardo Viana Vargas*, n.d.).

In addition, AI will have a significant impact on forecasts and estimation, automating processes such as the identification of patterns to better pinpoint projects with a higher success rate and value proposition or that are ready to launch. Moreover, it will ensure a more balanced portfolio with a lower risk. Apart from the superior computational power, the key behind this benefit is that AI removes biases that have been an everlasting plague of decision-making.

Lastly, testing processes will be impacted. To this day, testing is an essential part of projects and AI will improve testing and make it easier and more available even for smaller projects, allowing early detection of errors and issues and in some cases, automatically starting self-correcting processes, reducing the number of tests and reworks (Nieto-Rodriguez & Viana Vargas, 2023).

All these changes will give the possibility to PMs to shift their focus to soft skills, strategic thinking, and leadership. This will allow for a more accurate delivery of the expected benefits and strategic goals' alignment. Nevertheless, they will significantly change the shape of project management.

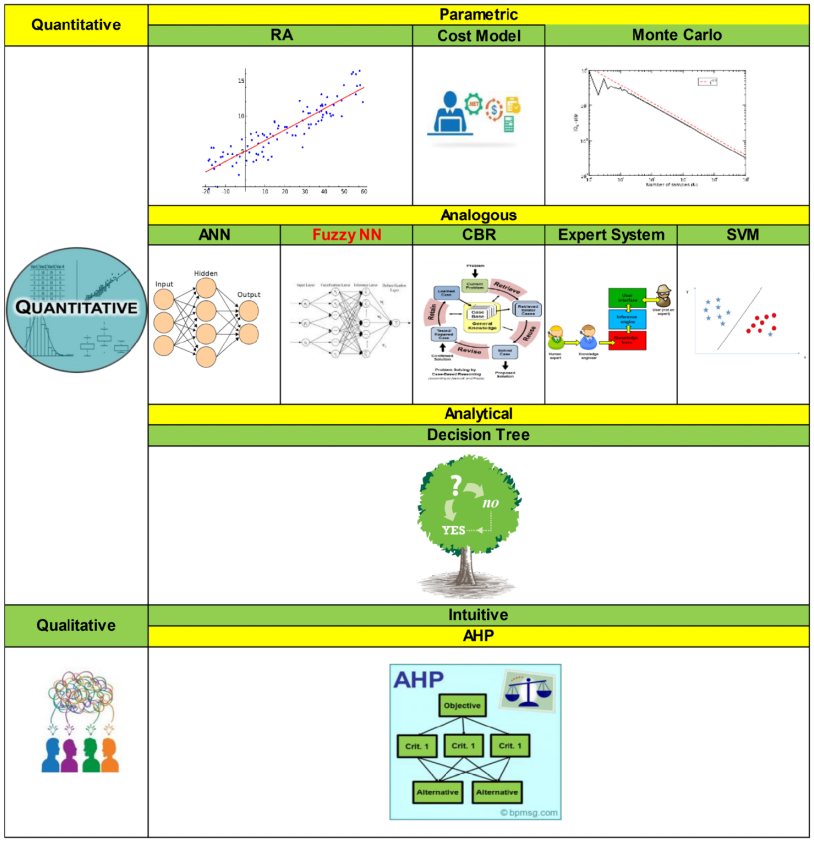
AI Techniques Categorization

With the advancement of AI and with the previously mentioned opportunity of increased estimation accuracy, the literature review naturally revealed intense research in this field. These several new AI-based techniques can be roughly divided by the technology they implement: Machine Learning Systems, Knowledge-Based Systems, and Evolutionary Systems. They all share the common trait of regarding the project as an organic whole, being able to include all the problem's features, briefly presented in the following chapter, to achieve an estimate where every element of the project and the interaction it has with the others, with the resulting consequences on the whole project, are accounted for.

It might be of interest to position these new techniques in the traditional division of qualitative and quantitative methods. In construction projects, quantitative estimating methods, the most employed for construction projects (Hashemi et al., 2020), can usually be subdivided into parametric, analytical, and analogous, while qualitative methods include statistical, intuitive, or expert judgment methods. The new computational approaches presented in this thesis can be allocated to the category of the traditional Analogous estimation, as represented in Figure 1.1.

Figure 1.1

Categorization of the innovative methods employed for cost estimation



Note The figure displays the different computational methods now employed for cost estimation and their allocation between qualitative and quantitative methods (Hashemi et al., 2020).

Methodology

When conducting the systematic review at the base of this thesis, it was imperative to ensure that only eligible publications were included. To achieve this, the author followed the PRISMA model (Page et al., 2021) for systematic reviews reporting, and considered all published research studies and publications regarding construction projects written in standard English. Unpublished material was excluded from the analysis to ensure that the review comprised only credible and reliable sources of information. Only studies with a clear description of the datasets and methodologies, along with a detailed specification of the algorithms and procedures adopted and the rationale behind their choice

of output and input variables were included in this review. Moreover, only full-text available studies were included.

The exhaustive list of studies and materials mentioned in this paper has been included in the Resources section, located at the end of this thesis. More specifically, the studies investigated and included during the systematic review have been reported in the Results section. The studies have been selected with the overall aim of providing an understandable overview of the techniques and systems presented. To achieve this objective the following studies were investigated but not presented: studies with complex hybrid models that were difficult to grasp from management and engineering students without a background in AI; models focused only on variables' selection and optimization; studies not focused specifically on cost estimation; and studies that lacked clarity.

The search for these materials was conducted through several academic search engines, including Google Scholar, ScienceDirect, and ResearchGate, utilizing the combination of relevant keywords - cost estimation AND construction projects AND AI. For more in-depth research for the chapters Machine Learning Systems, Knowledge-based Systems, and Evolutionary Systems, the keyword "AI" was substituted with the name of the chapter for instance, and then with the name of the specific algorithm belonging to the class under analysis. Moreover, some studies were identified from citation searching.

The screening and collection process, a crucial phase of the research, was conducted meticulously by the author of this document alone. With the utmost attention to detail, the materials were obtained using the search method described above, and the information was verified through the analysis of the resources cited and cross-referencing with other relevant studies already employed in the research, if available.

During the review, the primary objective was to provide a comparison between the accuracy of cost estimates generated with different AI-based techniques compared to each other and traditional methods.

The focus was not just limited to the technical aspects but also encompassed the state of diffusion of AI software in project management and its impact on critical issues. As a professional review, only verified information was included, leaving out any missing or unclear data, as the author had no means of verifying them.

The overall diagram for the systematic screening is reported in Figure A.1 of the Appendix.

The following biases have been identified: informational bias, observer bias, and cognitive biases such as confirmation bias and availability heuristics.

The informational bias, which is defined as "a distortion in the measure of association caused by a lack of accurate measurements of key study variables" (Alexander et al., n.d.), may have occurred by the selection of research studies with inaccurate measurements or input data set, which are not easily accessible. Moreover, this bias may have affected the studies cited and employed in this review, and thus, the results here displayed might present the same bias.

The observer bias is described as a systematic difference between the truth and the information observed and recorded during the study (Mahtani et al., 2018) and thus might have affected both this paper and the studies cited.

Confirmation bias, which the Encyclopedia Britannica defines as "people's tendency to process information by looking for, or interpreting, information that is consistent with their existing beliefs" (Casad & Luebering, 2023), might have affected the author's search for literature studies, disregarding articles with low accuracy rate for AI-based techniques, blinded by the desire to prove the improved accuracy of AI software for cost estimation.

Lastly, availability heuristic is "a common strategy for making judgments about likelihood of occurrence in which the individual bases such judgments on the salience of the information held in their memory about the particular type of event: The more available and relevant information there is, the more likely the event is judged to be" (*Availability Heuristic*, n.d.). This bias may have affected

the author with the preconceived idea of the lack of studies and data on AI algorithms adopted for pre-tender estimates in real projects, and not only as a follow-up study on historical data.

To mitigate the potential risks of biases and inaccuracies, a thorough review of the research papers was conducted, evaluating their risk assessments. Furthermore, the search and assessment methodologies employed were also carefully examined to ensure their validity and reliability. As an additional precautionary measure, three professors from different academic backgrounds, the supervisors listed on the first page, scrutinized the findings, limiting any potential sources of influence.

Given the nature of this review, a planned synthesis and sensitivity analysis was not considered appropriate or needed, as the review employs and presents literature studies, and does not include the design and application of algorithms to a common dataset to evaluate the different techniques' accuracy.

The confidence of the results was assessed by the consistency of findings across the literature studies under review with the same characteristics of algorithms and input variables. Unfortunately, due to the early stages of the research in this sector, and the limited number of studies, a significant number of different studies employing the same model over diverse datasets are not available to compare at the present day.

Costs, Estimates, and Factors in Construction Projects

The accuracy of cost estimation in a project is critical for the success of the project itself. When speaking about a construction project, the issue becomes particularly dire, as the data available is often insufficient and not accurate, but the costs are indeed a critical component of the project's success. Generally, cost estimations are conducted with one or a combination of these approaches:

Production Function

This function is modeled as a relationship between the volume of construction and a factor such as capital or labor. Indeed, the volume may be derived as a function of various input factors, employing mathematical and statistical methods. Thus, far a certain required volume the overall aim is to solve a cost minimization problem, finding the correct combination of input values (*Approaches to Cost Estimation*, 2018).

Empirical Cost Inference

This system requires statistical techniques to help establish a relationship between the cost of construction and the characteristics or attributes of a system. Empirical cost inference aims to estimate parameter values to assumed cost functions by using regression analysis techniques (*Approaches to Cost Estimation*, 2018).

Unit Costs for Bill of Quantities

This method of cost estimation prescribes the assignation of a value to every unit of labor and material in the bill of quantities to obtain the total cost as the sum of the products of the quantities multiplied by their unit cost. Although this methodology is time-consuming and complex, it generates a comprehensive breakdown of the various components of a project as well as the total cost, generating a thorough estimation (PMI, 2013).

Allocation of Joint Costs

By definition, a joint cost is an expense incurred in a joint process (Dan, 2013). Joint costs may include costs such as direct material and labor along with overheads generated during a joint

production process, in which an input automatically generates more than the desired outcome as byproducts. The principle behind the allocation of joint costs is that each expenditure item can be reconducted to a specific aspect of the operation. Virtually, the allocation of joint costs should be causally related to an original single costs category. Nevertheless, sometimes, a causal relationship between the allocation factor is pernicious to find.

Types of Costs in Construction Projects

The total cost of a construction project includes both the cost of construction and the cost of maintenance and operation (PMI, 2013). The initial cost estimate, called capital cost estimate, is of interest to the construction manager, and includes the expenses related to the initial establishment of the facility such as land acquisition; planning and feasibility studies; architectural and engineering design; construction materials, equipment, and labor; construction financing; field supervision; insurance and taxes; owner's general office overhead; equipment not included in construction; inspection and testing. Conversely, the operation and maintenance cost, purview of the client, includes elements such as land rent; operating staff; labor and material for maintenance and repairs; periodic renovations; utilities; financing costs; insurance, and taxes.

Depending on each project, different cost categories will have different impacts. Hence, according to every situation, certain classes must not be overlooked by estimators. At the same time, an accurate estimate of the operation and maintenance cost is paramount to conduct a robust life cycle cost assessment.

Another element of construction budgets is the contingency reserve, saved for unexpected possible cost generated by identified risks. It is calculated by accounting for every identified risk, with different techniques such as (Usmani, 2022):

-Percentage of the Project's Cost

-Expected Monetary Value

-Decision Tree Analysis

-Monte Carlo Simulation

The contingency sum for each item can be allocated to each cost unit or in a general category of construction contingency. Contingency amounts not spent for construction can be released near the end of construction to the owner or to add additional project elements.

Conversely, the Management Reserve is set aside to address unidentified risks. While, It is part of the budget, it is not accounted for in the cost baseline, and it is usually calculated following organizational directives (Usmani, 2022).

Cost Estimates' Categories in Construction Projects

Different organizations recommend various methods of grouping cost estimates' categories. Nevertheless, there are two main methods of grouping cost estimates. The classification in five different categories, according to the scope and level of accuracy, is analyzed in the following pages, and the more streamlined process of clustering the costs into three main macro-categories is presented after the five classes' method.

Class 5 Estimates or Rough Order of Magnitude Estimates

Class 5 estimates are used in a series of strategic business planning processes, such as market studies and project screenings. These estimates are hence calculated in a limited amount of time and with little information, such as only the building type, location and functional space required, and total building area. This information accounts for 0 to 2% of the project definition (Christensen et al., 1997).

It is then understandable why the accuracy of these estimates ranges from -30% on the low side and +50% on the high side. For these types of estimates, stochastic estimating methods are usually employed, such as parametric techniques and modeling techniques such as cost/capacity curves, the rule of six-tenths, the Lang factor method, the scale of operation factors, and cost indices (Hendrickson, 2008).

The poor accuracy of these estimates poses an often-overlooked bias issue. As Kahneman and Lovallo presented in their paper '*Delusions of Success*', cognitive biases have a significant influence on pre-charter estimates (Lovallo & Kahneman, 2003). Biases are a cardinal flaw in decision-making given by the human brain. There are two main types of biases: optimism bias, which leads individuals to perceive themselves as superior to others in their performance, and planning fallacy, which occurs when individuals assume they won't encounter the same obstacles others face. Optimism is further amplified by other biases: anchoring, competitors' neglect, and organizational pressure. The first one is the most pernicious, as it happens when the human brain latches on to a set of numbers and keeps them in mind. Their influence will skew any following hypothesis, even if there is no correlation between the two. The second one can be quite disruptive in situations such as price wars, as, in making forecasts, estimators tend to focus on their company's capabilities and are thus prone to neglect the potential abilities and actions of rivals. The last one is a conscious modification of the estimates to increase the selection chances of the proposed project.

Class 4 Estimates or Feasibility Study Estimates

Class 4 estimates are employed for concept evaluation, feasibility evaluation, and preliminary budget approval by organization heads and construction managers for strategic business planning. They involve an increased number of resources and amount of detail as they will determine the viability and economic value of a project, along with a means to evaluate possible alternatives. At this stage, from 1 to 15% of the project deliverables need to be defined (Christensen et al., 1997), and elements such as preliminary room layouts, site plans, drawing markups for demolition and utilities, and technical memorandums need to be defined. The typical accuracy ranges from -20% to +30%. The cost estimation methods used are the same as those used in Class 5, although they are applied to specific items (Hendrickson, 2008).

Class 3 Estimates or Budget Estimates

These estimates are prepared for budget authorization and funding. Thus, they usually constitute the initial control estimate employed to monitor costs and they are calculated from a preliminary engineering design. The engineering phase of the project at this stage is 10% to 40% completed (Christensen et al., 1997). Indeed, the building code or standards requirements; the exterior closure description; and the finishes descriptions and requirements are all defined. In many projects, these estimates are the most detailed ones and are the only basis for cost and schedule monitoring, even if Class 3 cost estimate accuracy ranges from -15% to +20% (Hendrickson, 200). These estimates generally involve deterministic estimating methods, employing a high degree of unit cost line items, which can be at an assembly level of detail. Factoring and other stochastic methods might still be employed to estimate less critical areas.

Class 2 Estimates or Bid or Tender Estimates

These estimates often constitute a contractor control baseline implemented to monitor and control costs and progress. Contractors also employ them as bid estimates. For these estimates, the project already sees the majority of the deliverables defined, such as draft specifications, building systems, and soil and hydrology reports. Drawings and datasheets of process units, equipment, and utilities are completed. Additionally, the project execution plans are finalized, with the WBS, timeline, equipment, manpower and material scheduling. Overall, 30 to 70% of the deliverables are determined (Christensen et al., 1997)(Hendrickson, 2008). The estimating method used involves tens of thousands of unit cost line items instead of factoring methods seen from the previous estimate classes. For undefined areas, an assumed level of forced detail is calculated. The accuracy range goes thus from -15% to +20%.

Class 1 Estimates or Full Detail Estimates

These estimates are conducted for single aspects of the project, usually for subcontracting purposes. They are viewed as current control estimates, used to support the change management

process. They may be used to evaluate bid checking, support vendor/contractor negotiations, or for claim evaluations and dispute resolution (Borowicz et al., 2020). Typically, engineering is from 70% to 100% complete (Christensen et al., 1997) and would comprise virtually all engineering and design documentation of the project, and complete project execution and commissioning plans (Hendrickson, 2008). Class 1 estimates are accurate from -5 to +10%. They are very detailed and thus are calculated with deterministic estimating methods on critical aspects with unit cost line items (Borowicz et al., 2020).

As mentioned above, some organizations prefer to use a more streamlined approach grouping the estimates into three macro-categories with fewer levels of distinction (Hendrickson, 2008).

Design Estimates

They reflect the progress of the design during the initial stage, the pre-design, and the detailed design phases of a project. This category groups together estimates with progressive levels of details from Class 5 to 3, following the development of the project. Indeed, this category encompasses (PMI, 2013):

-Screening estimates (or order of magnitude estimates)

-Preliminary estimates (or conceptual estimates)

-Detailed estimates (or definitive estimates)

-Engineer's estimates based on plans and specifications.

Bid Estimates

A bid estimate may be prepared by the contractor or subcontractor, and it is based on the list of all the materials necessary for the project or the construction procedures chosen by the contractor. Thus, it corresponds to Class 2 estimates. Indeed, the direct cost of construction for bid estimates is usually derived from a combination of the following approaches (PMI, 2013):

-Subcontractor quotations

-Quantity takeoffs

-Construction procedures.

Control Estimates

These estimates are used by both the owner and the contractor, as they will need to establish a budget to either plan the long-term financing of the project or use it for cost control, respectively. Specifically, the owner will use the most detailed estimate, which could be according to the organization a class 3, 2, or 1 estimate, and the contractor the bid estimate. Both will need to be periodically revised to reflect an accurate cost to completion. For monitoring the project during construction, a control estimate is derived from available information to establish (PMI, 2013):

-Budget estimate for financing

-Budgeted cost after contracting but before construction

-Estimated cost to completion during the progress of construction

Variable Factors in Construction Projects

In general, the accuracy of a construction cost estimate is directly connected with the level of detailed information, from building dimension to position (Kim et al., 2004). Unfortunately, the construction sector has not yet widely established the systematic practice of documenting and publishing project data with public access (Mohamed & Moselhi, 2022).

As it is possible to imagine, the factors affecting a project differ for each unique case, and every industry has its historical data. Additionally, it is important to remember that, to ensure quality, durability and robustness, other parameters such as the shear strength of soil and compressive strength of concrete ought to be estimated, which may not be directly linked to the project budget but do nevertheless influence the overall performance (Sharma et al., 2021). In fact, since it is well known that inaccurate estimation of quality parameters leads to cost overrun, delay in completion, and damage to structures, their estimation is particularly crucial. Nevertheless, early prediction and estimation of these project parameters is a challenging task as the preliminary phases of a project are characterized by little information available. Manual calculation is still widely popular, impinging on

the project by increasing cost and time. With artificial intelligence, estimations could be conducted automatically, with the benefits of cost and time efficiency correlated by significantly increased accuracy and the loss of human error. By employing this approach, the estimation of pre-construction outputs such as the ones previously mentioned, could be conducted despite the lack of an abundance of parameters, ensuring a better delivery of the project (Sharma et al, 2021).

As mentioned above, there are various factors and unexpected variables, which vary from project to project and can be both internal and external, that significantly impinge the calculation of the parameters. Given the fact that they represent the construction's features and impact the overall cost, there must be a correct enumeration of these factors to best represent the construction's features. One may think that to achieve better accuracy, an increased number of factors is needed. However, increasing the number of factors leads to an increased number of estimation cases. Additionally, there is no evidence that an increased amount of input data leads to an increase in the accuracy of estimation, thus the correct number of factors that best represent the construction's features are needed to increase the estimate's accuracy (Sharma et al., 2021). Following an extensive literature review it can be summarized that, generally, the most crucial parameters such as the type of terrain and location, project and labor size, are usually evaluated. Moreover, in most of the studies, experts in the fields and consultants, along with previous literature and historical data, are usually questioned to define the type of input. Additionally, there are other important factors, which can be grouped as design and project-specific factors, presented ahead that, given their significant influence over the project, are usually included in input factors of prediction models.

Project Complexity

The International Centre for Complex Project Management (ICCPM), states that "complex projects are characterized by uncertainty, ambiguity, dynamic interfaces, and significant political or external influences" (*What Is Complexity in Project Management?*, n.d.). Several research papers indeed reveal different complexity dimensions, constituting variables or factors that influence each

other and have a final effect on costs and durations (Herszon & Keraminiyage, 2014). Various methodologies and solutions are presented to overcome this issue, and there is clear evidence, analyzed in the following chapters, that with AI, projects' complexity is easier and more feasible to address.

Clear Specification

Thorough specifications of the requirements can prevent subsequent changes to the scope, which leads to increases in costs and duration. Oftentimes, teams struggle with developing clear requirement statements when the stakeholders' involvement is not assured and stable. Moreover, several issues might arise with the involvement of stakeholders. For instance, not all the stakeholders are always acknowledged, an error that stems from an insufficient and approximate analysis of the projects' overall effects (Belack, 2022). Secondly, stakeholders' requirements are routed throughout the organization to gain alignment and buy-in. Thus, often the requirements are not uniformly understood and do not represent the full breadth of the project needs.

Literature is rich in tools and methodologies to help teams manage stakeholders, such as Joint Application Design (JAD). This facilitates meetings to effectively improve the requirements definition activity by incorporating collaborative meetings and group management techniques led by a neutral facilitator (Burek, 2008).

Clear Scope Definition

This factor is closely related to the clear definition of specifications and requirements. A clear definition of the scope is the cornerstone of every project. Indeed, in the scope definition, the grounding elements of the entire project are defined and most of the issues which might arise later on are related to a poorly defined aspect of the scope (Greiman, 2013). The project team and estimators should, therefore, remove any uncertainty in the scope and make it transparent and universally understandable. Indeed, it must be clear, with the definition of the activities, resources, constraints, and deliverables. The team needs to clearly state the included and not-included deliverables, along

with the acceptance criteria and procedures to avoid scope creep. Lastly, all possible ambiguities, risks, and assumptions need to be addressed.

Prior Experience of the Contractor and Staff for Cost Estimation

Ideally, both the estimation team and the contractor should have extensive prior experience, but this is not always feasible. Cost estimates should be prepared by a multidisciplinary team with functional skills in financial management, engineering, acquisition, logistics, scheduling, mathematics, and communications, along with participants from all the parties majorly affected by the estimate. According to the U.S. Government Accountability Office (GAO) "a cost analyst should possess a variety of skills to develop a high-quality cost estimate that satisfies the 12 steps of a reliable cost estimate" (U.S. Government Accountability Office, n.d.). Those are: economics and accounting to properly address the inflation effects and the accounting systems; budgeting for properly allocating the funds over time; awareness of engineering, computer science, mathematics, and statistics will help identify cost drivers and the type of data needed to develop the estimate. The estimator should also possess adequate technical knowledge when meeting with functional experts to establish credibility and a common understanding of the technical aspects. In addition, cost estimators need good communication and presentation skills to deliver a convincing and solid presentation and to properly communicate with stakeholders to identify the project's specifications (U.S. Government Accountability Office, n.d.).

Equipment Requirements

The impact of equipment on a project's costs and duration is considerably relevant. Any change to the equipment may affect the overall cost and duration of the project. Indeed, equipment costs can be as much as 50% of an equipment-intensive job like construction projects. There is however an important difference to acknowledge between job cost and equipment cost.

While job cost can be used to monitor and control construction while in progress, equipment cost information provides "the long-term operating cost and the respective performance of individual

pieces of equipment as profit centers unto themselves" (Equipment Cost Setup, n.d.). Historical equipment costs are crucial to providing data for better accuracy. Without equipment costs, the operating cost in the bid will not be precise, leading to an inaccurate estimation. Moreover, the contractor will have to use the market rental standard rate for the equipment, failing to consider the cost advantage of in-house equipment. Hence, with equipment cost data available cost estimation is increasingly accurate (Equipment Cost Setup, n.d.).

Consideration of Site Constraints

The site is critical to the project; thus, its constraints should be fully analyzed for cost elements that are unique and present an impact on the overall costs (Sharma et al., 2021). Indeed, the literature is rich in projects where costs overrun due to inaccurate site evaluations.

Availability of and Consultation with Previous Similar Bids

Historical data are fundamental when generating new estimates. Truly, one of the approaches proposed to overcome estimating biases by Kahneman and Lovallo is incorporating in the planning processes an objective forecast method, such as reference-class forecasting. It entails disregarding the details of the project at hand, examining the experiences of a class of similar projects, drafting a rough distribution of outcomes for this reference class, and then positioning the current project in that distribution. Literature research proved that this method leads to more accurate results (Lovallo & Kahneman, 2003).

Change in Overall External Environment

This factor is crucial for long-term projects. Indeed, despite the best forecasts, it may be difficult to foresee the technological, societal, political, and economic changes that will occur throughout the project.

One glaring example of the importance of this factor is the Big Dig project (Greiman, 2013). Several external influences impacted the Big Dig project of Boston, especially given its long duration, resulting in severe delays and cost increases. The Central Artery/Tunnel Project, also known as the

Big Dig project, was the largest, most challenging highway project in the history of the United States. The project aim was to replace Boston's six-lane elevated Central Artery (I-93), with an underground highway and two new bridges over the Charles River. It also extended I-90 to Boston's Logan International Airport and Route 1A. When planning for the CA/T Project began in 1982, nobody could have predicted the challenges that lay ahead. Congress approved federal funding and the project's basic scope in April 1987 (*The Big Dig: Project Background*, n.d.). The Big Dig's original budget was set to $2.5 billion, with an overall duration of 9 years. However, the costs skyrocketed to $14.8 billion and was not finished until 2006. This incredible undertaking employed 5000 workers, 130 major contractors, and a large fleet of equipment including more than 150 cranes. The excavation dirt was enough to fill a football stadium 16 times. Indeed, after the project the Boston Harbor sported a brand-new island created with the excavation material (Greiman, 2013).

With such a long duration and massive size, this project encountered several problems due to the changing external circumstances. Some of the most impactful ones were regulation and political changes, technology changes, environmental factors such as historic preservation requirements, bridge foundations, contaminated soil, seasonal restrictions, natural hazards, market factors and inflation, community concerns, and traffic demands. For example, a relevant amount of funding came from the federal government and along the life cycle of the project, any changes in the regime of the government impinged upon the amount of funding received, both positively and negatively.

To manage these influences, literature papers now recommend the use of Scope Change Management systems. Specifically, to manage the schedule's changes, it is important to keep stakeholders involved from the conception of the project, using incentive systems to keep them motivated and be sure they meet the requirements and communicate with the community via an efficient PR office. To manage cost increases, it could be useful to develop fully integrated cost and resource project schedules to provide greater understanding to managers and the public in tracking and managing the whole project, asking for consultancy from experts on the requested changes.

Additionally, to minimize the effects of inflation it is important to periodically update the cost escalation and try to anticipate cost increases (Belack, 2022).

Number of Competitors

According to various studies, it was observed that increases in the level of competition lead to excessive cost overrun. The number of competitors can be computed as the total number of bidders who file their bids for a project. Bidders often quote unrealistic values for a project to achieve the lowest bid for the project, leading to cost overrun during the project at later stages (Hyari et al., 2016).

Awarding the Contract to the Lowest Bidder

Owners generally grant project contracts to the lowest bidders; however, these might not always be the best choice. Request for Proposals (RFPs) help ensure transparency and show the public they're accountable for their project goals and vendor choices. Additionally, RFPs allow all vendors to bid on their projects (Jordan, 2021). Unfortunately, to win the bid, the estimates might be not realistic, leading to poor results and delays, thus causing an overall cost and duration increase.

Pre-qualification criteria and policies followed when granting the project need to be strengthened to prevent the problem (Sharma et al., 2021). For example, in governmental RFP, the award is specified to go to the "lowest responsive and responsible bidder" according to the NIGP. The responsive contractor is defined as "a contractor, business entity, or individual who has submitted a bid or proposal that fully conforms in all material respects to the Invitation for Bids (IFB)/Request for Proposals (RFP) and all of its requirements, including all form and substance" and a responsible bidder as "A business entity or individual who has the financial and technical capacity to perform the requirements of the solicitation and subsequent contract. When combined the winning bidder is defined as "the Bidder who fully complied with all of the bid requirements and whose past performance, reputation, and financial capability is deemed acceptable, and who has offered the most advantageous pricing or cost-benefit, based on the criteria stipulated in the bid documents" evaluating other factors apart from price (NIGPT-The institute for public procurement, n.d.).

Frequent Changes in Design Specification

Frequent changes in the design specification of the project usually lead to wastage of time and material and subsequent cost overrun. Different types of clients may be more or less invested in the design phase. In some cases, a design firm can be appointed to create the drawings and inspect the project phases. Regardless of the level of involvement of the client in the project, the scope must be clearly defined in the pre-charter phase (Greiman, 2013).

Indeed, clear guidelines for managing scope change and the authority required for it must be included in the statements. As explained in previous paragraphs, it is imperative to engage external participants and assess the environmental conditions that can influence project costs. The loss of scope control, particularly during engineering, is one of the major factors leading to discrepancies between estimated and actual costs. This complication oftentimes is the product of a few major changes to the scope or of successive minor changes, often referred to as scope creep.

A clear specification statement, whose importance was analyzed above, allows for immediate realization when extra work is added and is a beneficial tool for controlling scope creep. Nevertheless, it is sometimes not enough to prevent loss of control, especially if a process for scope control is not followed, or if the scope statement is not precise and comprehensive of all the stakeholders' requirements. A way to prevent scope creep is to follow a specification management plan, which indicates how scope changes will be identified, how they will be integrated into the project, what approval requirements are needed, and from whom. Specification management controls may also include progress performance requirements to enable stakeholders to understand how frequently the specification might change and by how much (Greiman, 2013).

Material Costs and Their Fluctuation

Recent crises have highlighted the instability of the global supply chain and, the unavailability and great price fluctuations of materials. Economists and industry professionals have identified numerous factors behind these fluctuations. Material price fluctuations constitute an ever-present and

significant risk to construction projects. These fluctuations cause both unforeseen cost increases along delays. The major solutions presented by consultants and experts are contractual provisions; and taking proactive measures (Pray & Walsh, n.d.).

Contractual Provisions. Contractual provisions have been employed for decades to address exchange-rate fluctuations. These diverse provisions allocate the risk of material cost increase with different rates to different parties, according to the situation. Apart from risk allocation, price fluctuations could also constitute grounds for suspension or termination of the contract. Owners might also employ them as cause for suspension of the project (Pray & Walsh, n.d.). This option can be useful if the unavailability of materials impacts the critical path, causing delays and costs.

Proactive Measures. Even if the contract does not address price fluctuations, it is possible to adopt proactive measures to lessen their impact. The key element for reducing these impacts is communication. If suppliers do not promptly communicate price impacts and product unavailability to the general contractor, and the latter does not communicate them to the owner, it will be quite challenging to address the issues (Pray & Walsh, n.d.). Several options are grouped in these categories, such as identifying alternative suppliers, direct purchasing to reduce markup, purchasing the material needed during the early stages of the project which requires warehouse space, and defining alternatives to the material specified in the original design.

All the elements presented and analyzed above, as already mentioned, are crucial to define the cost estimating problem, as they describe the construction project under analysis and allow for increased accuracy of the estimating output.

Machine Learning Systems

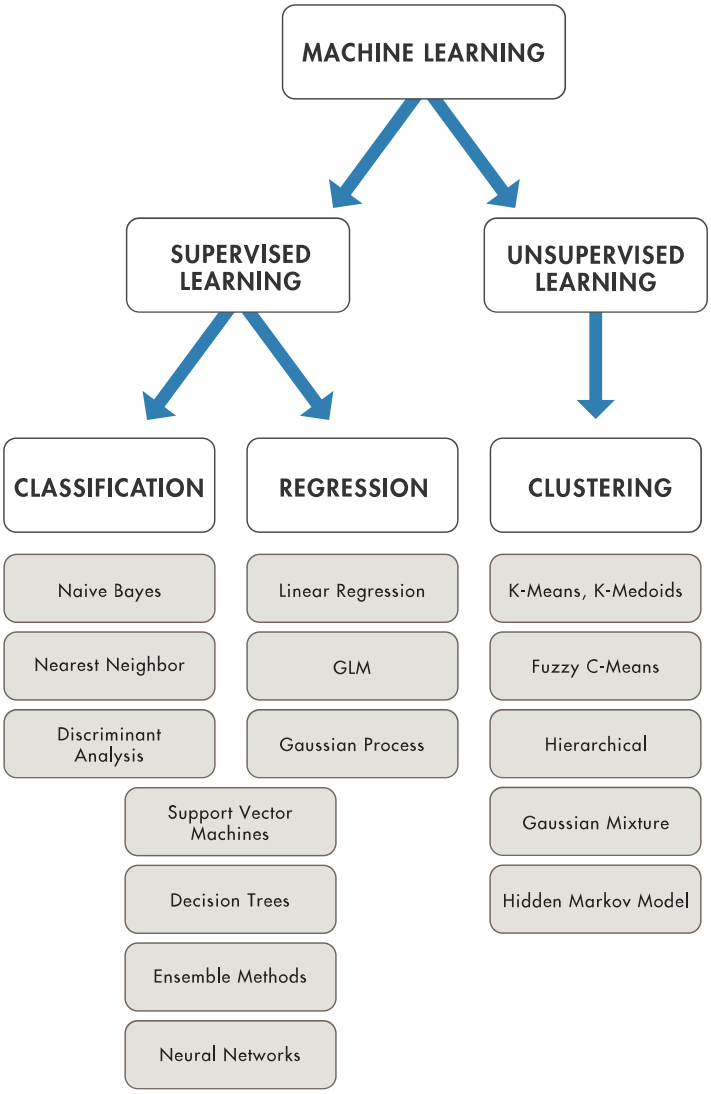
According to IBM "Machine learning is a branch of artificial intelligence and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy" (What Is Machine Learning?, n.d.). This science, born in 1950 with the Turing test, consists of coding a program that learns to perform a task from experience, and with it, it improves its performance (Géron, 2019). It now enables speech recognition with ChatGPT (OpenAI, 2023), the identification of terrorist suspects and to autonomously drive cars, amongst other things (Marr, 2016). Machine Learning Systems (MLs) are particularly useful for situations in which they can simplify the code in case of long lists of rules or fluctuating environments, as they can deal with uncertainty and incomplete data. Unfortunately, the logic and justification that motivate the final solution obtained by these algorithms are not open to scrutiny, as it is a black-box decision. Nevertheless, they are particularly useful for data mining and pattern recognition (Géron, 2019).

Machine Learning can be subdivided into different classes with different classification criteria, reflecting different characteristics of the algorithms. The most adopted classification categorizes the algorithms according to the level of human supervision required: Supervised Learning, Unsupervised Learning, and Reinforcement Learning. The first one trains a model on known labeled input and output data so that it can predict future outputs; the second finds patterns in unlabeled input data; and the third one is focused on providing feedback to the algorithm after each iteration to evaluate the choice made[1]. A schematic view of the inner classification of ML is depicted in Figure 3.1.

[1] For further information on Reinforcement Learning please refer to '*Hands-On Machine Learning*' (Géron, 2019).

Figure 3.1

Methods of Machine Learning



Note The figure depicts the algorithms belonging to machine learning and their classifications (What Is Machine Learning?, n.d.).

Supervised Machine Learning

According to MatLab, "this method builds a model that makes predictions based on evidence in the presence of uncertainty" (What Is Machine Learning? n.d.). It is utilized when there is a predefined, labeled, set of input data and corresponding known output, to train a model to make

accurate predictions for new information, much like Spam filters, where users flag spam emails and the system classifies as Spam similar emails (Géron, 2019). Supervised learning algorithms can typically perform Classification and/or Regression tasks.

Classification Techniques

These algorithms organize input into categories and thus are employed with discrete output. They are employed in medical imaging, speech recognition, credit scoring, and other cases. Examples of algorithms belonging to this category are Decision trees (DT), Random Forest (RF), and Support Vector Machine (SVM) (Machine Learning Algorithms, n.d.).

Decision Trees. They are utilized for classification and regression tasks alike. As the name suggests, they present a tree-like structure, with a root node, branches, internal nodes, and leaf nodes, which constitute the possible outcomes in the given dataset. The key objective for the algorithm, in case of classification for example, is to identify the best split points. The process of splitting is reiterated continuously until all records have been classified under specific class labels. While these algorithms are white box approaches, with clear decisions and little preparation for the data set (Géron, 2019), they do present the issue of purity, in the sense that leaf nodes aim to represent a homogenous set of points, which can prove increasingly difficult when the complexity increases. Hence, up to a certain size, decision trees are easily able to clearly group points in a single class, then this clear separation becomes less evident and data points do not clearly belong to a single leaf node (*What Is a Decision Tree*, n.d.). This occurrence can lead to overfitting, which means that the algorithms have attuned to the noise of the dataset and perfectly fit the input data, being thus unable to manage new data on which was not trained (*What Is Overfitting?*, n.d.). To prevent overfitting, complexity can be reduced by the regularization of hyperparameters[2], which in this case means tuning

[2] Hyperparameters are regularization parameters of the learning algorithm, set before training (Géron, 2019).

the depth of the tree (Géron, 2019), reducing the number of nodes on non-relevant features, or employing a random forest algorithm.

Random Forest. The Random Forest algorithm (RF) is an alternative to Decision Trees that overcomes the issue of overfitting and presents more flexibility. However, it is more time-consuming and complex to interpret along with a bigger dataset (*What Is Random Forest?*, n.d.). This algorithm is an example of the ensemble method training with the bagging method. The former sees a random sample of data in a training set selected with replacement several times to create different samples, trained separately so that the average of their results will be the final result. Conversely, the latter sees the algorithm trained on different subsets of the training dataset, whose sampling is performed with replacement. Hence, Random Forest is "a group of Decision Tree classifiers each on a different random subset of the training set" (Géron, 2019). As previously mentioned, it is quite often employed to avoid the overfitting issue that could impinge on Decision trees, as they select the best node-splitting feature over a random subset of these, introducing more diversity (Géron, 2019).

Support Vector Machine (SVM). Support Vector Machine is an algorithm employed to find "a hyperplane in an N-dimensional space that distinctly classifies the data points" (Gandhi, 2018), and it is adopted for either classification or regression means. As hyperplanes are decision boundaries (with a number of dimensions equal to the one of the spaces that represent the number of features) that help divide data points into different classes, the ultimate aim is to find the hyperplane with the maximum margin (distance between the points belonging to different classes). To achieve this result, data points that are closer to the hyperplane, called Support Vectors, are identified, and employed to maximize the margin, through maximizing a loss function.

Regression Analysis

Regression Analysis (RA) is employed to predict continuous responses, such as prices of financial assets. This method is utilized for studying the relationships between different variables and forecasting a dependent variable based on its relationship with independent ones. At a visual level,

this process translates into estimating a regression line from a group of data, usually represented as scatter plots. There are two main types of regression analysis: single variable linear regression, which has one independent variable and one independent, and polynomial regression, with three or more dependent variables. The equations that model these lines have the traditional polynomial form with the Y-intercept, slope, and error term. Statistical programs such as Microsoft Excel are typically employed to run regressions. The resulting equation, built on the input historical data, is employed to conduct the prediction of the dependent variable based on specific values of the independent variable. This approach is widely used nowadays in any application of data analytics. Indeed, this is a practical and accessible approach to obtaining predictions, which is pivotal for strategic decision-making (Cote, 2021). Conversely, another type of regression, logistic regression, employs the logistic function and can also be used for classification purposes. It is commonly employed as a binary classifier to establish the belonging of an item to a particular category (Géron, 2019).

Unsupervised Machine Learning

Unsupervised learning, conversely, employs unlabeled data and is implemented for tasks such as clustering, association rules learning, and dimensionality reduction.

Clustering

The mechanism behind this technique is to group data into different categories according to their similarities or differences. These algorithms are usually employed to process unclassified data into groups. Clustering algorithms can be categorized into exclusive, overlapping, hierarchical, and probabilistic.

Exclusive clustering is a form of grouping that stipulates a data point can belong only to one cluster.

In overlapping clusters points can belong to multiple clusters with different degrees of membership.

Hierarchical clustering can be Agglomerative or Divisive. Agglomerative Clustering is considered a bottom-up approach. Its data points are isolated as separate groupings initially, and then they are merged iteratively based on similarity, which can be evaluated with different methods based on the distance between the points until one cluster has been achieved. Divisive Clustering is the opposite and adopts a top-down approach. In this case, a single data cluster is divided based on the differences between data points.

In probabilistic clustering, data is divided on the likelihood that it belongs to a certain distribution (What Is Machine Learning?, n.d.).

One example of a clustering algorithm is the K-Means.

Association Rules Learning

The methods employed for this task aim to find relations between variables in a given dataset. They are commonly used for market basket analysis to understand the cross-influence between different products. They are indeed employed to determine bundles and selling strategies.

Dimensionality Reduction

Dimensionality Reduction is a technique that addresses the issue of an excessive number of dimensions in a dataset. It may combine features together to reduce the size of the dataset without impinging on its quality. It is commonly used in the training phase, with algorithms such as autoencoders, which are unsupervised artificial networks (Géron, 2019).

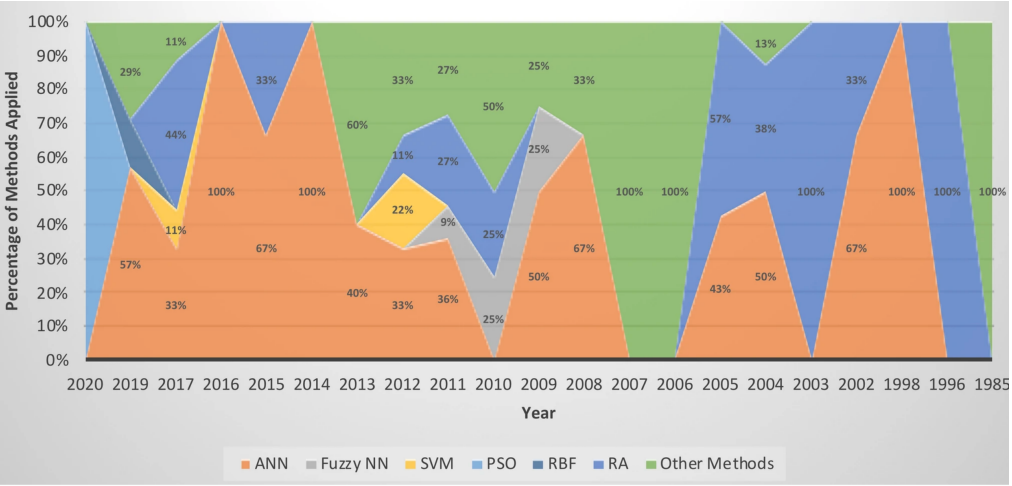
Machine Learning for Cost Estimation in Construction Projects

Following a literature analysis, it is possible to notice that the most used ML algorithms in the cost estimation field for construction projects are Artificial Neural Networks, which can belong either to supervised or unsupervised learning. Specifically, a recent review reported a comparison of the

popularity of different ML techniques: ANN, Fuzzy NN, PSO[3], SVM, RBF, and RA, displayed in Figure 3.2 (Hashemi et al., 2020). From the diagram, it is possible to notice that over the years, ANNs have been incrementally more popular, while RA's presence reduced. This can be justified by the increased computational power and the development of datasets large enough to sustain the training of such complex algorithms (Géron, 2019).

Figure 3.2

ML algorithms and their use rate over the years



Note The graph depicts the employment rate of the main ML algorithms in research paper studies on construction projects since 1985 (Hashemi et al., 2020).

Artificial Neural Networks

Artificial Neural Networks (ANNs) are a simulated representation of the brain neural networks. They are used for pattern recognition, clustering, and forecasting, as they are self-learning algorithms that analyze significant amounts of data at a high velocity (Hegazy, 2002). One of the

[3] Particle Swarm Optimization, a stochastic optimization method created by Dr. Eberhart and Dr. Kennedy in 1995, draws inspiration from the collective behavior observed in birds or schools of fish. It involves a population of particles, each representing a potential solution, navigating a search space in pursuit of the global minimum. While the exact location of the global minimum remains unknown to each particle, they all possess fitness values determined by the optimization's fitness function (Thevenot, 2020). Given the relatively small incidence of use of this method and its complexity, this paper will not address it in more depth.

most famous ANNs is the Google search algorithm. In the cost estimation case, the purpose of ANNs is to map the multidimensional space of cost predictors, which are all the factors that impinge on the costs of a project, into a one-dimensional space of construction costs, the final estimate. In the statistical sense, the problem comes down to solving a regression problem and estimating a relationship between the cost predictors being independent variables and the overall project cost being dependent variable (Juszczyk et al., 2018).

Modern ANNs are composed of layers of nodes: an input layer, one or more hidden layers, and an output layer. While there are different models of ANNs, from McCulloch and Pitts' neurons for logical computation to Perceptrons (Géron, 2019), in the majority of studies in the cost estimation field, the model adopted is the Multilayer Perceptron (MLP). Each node, connected to others belonging to the same layer and the layer below, is a threshold logic unit (TLU), which calculates the weighted sum of input data and then applies an activation function, needed to introduce non-linearity between the neurons. There are several activation functions: the identity, the Heaviside step function, the sign function, the logistic function, the hyperbolic tangent, or the rectified linear unit function with its several variants (Géron, 2019). When the input data have been weighted and summed, the output is compared to the threshold: if the threshold is surpassed, the node is activated, and the output data is sent to the nodes in the following layer as input for those nodes. The majority of ANNs are feedforward, so they move from the input to the output, in a single direction. Nevertheless, there are backpropagation models that move from the output to the input. These models are used to calculate the error associated with every neuron, to achieve a better fit for the parameters. This corresponds to the testing phase, where the accuracy of the prediction of the algorithm is calculated, usually with a cost or loss function such as MSE[4] (*What Are Neural Networks?*, n.d.).

---

[4] $MSE = \frac{1}{2m}\sum_{i=1}^{m}(\check{y} - y)^2$ Where MSE=Mean Square Error, i=index of the sample, $\check{y}$=predicted outcome,

There are several specifications of ANNs; however, it is of interest to introduce two algorithms that will be illustrated deeply in the two estimating algorithms presented later.

Fuzzy neural networks combine Fuzzy logic and Neural Networks, consequently sporting the learning skills of neural networks and the ability of Fuzzy logic to manage noise. In their simplest form, a fuzzy neural network can be viewed as a three-layer network, with a fuzzy input layer, a hidden layer containing the fuzzy rules, which are If-Then statements that represent the knowledge of human programmers about the process (*Fuzzy Rules*, n.d.), and a final fuzzy output layer (Nauck & Kruse, 1999).

Radial basis function networks (RBF) are neural networks based on the Gaussian RBF; a similarity function implemented as an activation function that is introduced to handle nonlinear datasets. These networks present a faster learning speed and are an efficient tool for both pattern recognition and function approximation. They are a feed-forward neural network with the hidden layer comprising two layers of network nodes. The weighted sum of the hidden units is then sent to the output unit (Behera et al., 2023). This allows the algorithm to transform the input into a linearly separable space, often outputting a space with an increased number of dimensions, based on Cover's theorem on the separability of patterns (Ramadhan, 2021).

Overall, there are numerous advantages to employing ANNs: they can develop forecasts with less statistical training and, as already stated, their ability to detect relationships and patterns between variables is significantly superior. Nevertheless, their biggest disadvantage is the fact that they essentially work with a black box mechanism, and additionally, they are not suited for every type of problem and require high computational resources (Hsu et al., 1995).

---

y=the actual value,
m=number of samples.

These algorithms have been widely studied in literature and, specifically, there are plenty of research papers presenting hybrid models combining an ANN and a meta-heuristic method to perform more accurate predictions (Bai et al., 2014). One example is the model of backpropagation neural networks and genetic algorithms that lead to more accurate predictions (Tkáč & Verner, 2016). There is ample research on the application of ANN-based techniques to civil engineering projects. Research also shows that ANN, as already mentioned above, in Figure 3.2, is, for now, the most popular method (Hashemi et al., 2020).

A significant study employing ANN for cost estimation in civil engineering is the study conducted by Wilmot and Mei on the estimation of cost escalation of highway construction projects in Louisiana for the 1998-2015 period. According to the authors, cost escalation is one of the most overlooked elements of an accurate cost estimation, although it is significantly impactful (Wilmot & Mei, 2005). Before their contribution, cost escalation was measured with overall fixed indexes such as the Engineering News Record Construction Cost Index and Building Cost Index. This method does not allow for the change of initial conditions, as it compares with a base year, 1996, how much it costs to purchase a package of a certain amount of man hours, plus a certain amount of fabricated standard structural steel, Portland cement, and lumber, with amounts and prices that can vary between the two different indexes and with the city of reference (*Using ENR Indexes*, n.d.). There are other methods, such as multivariate linear regression analysis, that consider the relationship between costs and influencing factors. Nevertheless, even these methods assume that an unchaining specific mathematical formulation may not always fit the data, and they overlook the correlation between variables such as the cost of labor, equipment, and materials.

Conversely, Neural Networks do not assume an implicit function and thus are more flexible. The model proposed by Wilmot and Mei uses a modified Federal Highway Administrations' Composite Bid Price Index, renamed by the authors Louisiana Highway Construction Index (LHCI), which presents this structure:

$$LHCI_n = \frac{\sum_{i=1}^{5} \bar{P}_{i,n} \cdot Q_i}{\sum_{i=1}^{5} \bar{P}_{i,1987} \cdot Q_i} \cdot 100$$

Where $LHCI_n$ = Louisiana Highway Construction Index for year n.

$\bar{P}_{i,n}$= average price of representative pay item i in year n.

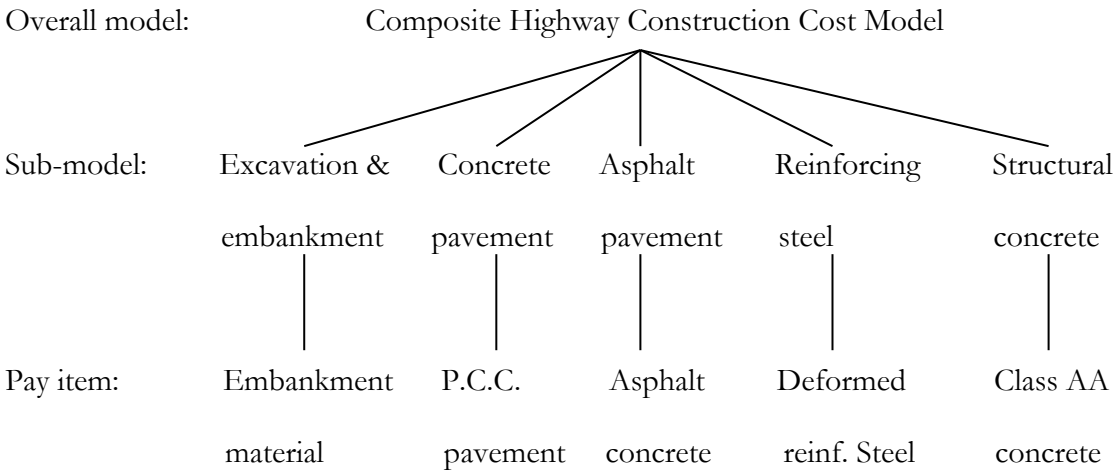
$\bar{P}_{i,1987}$= average price of representative pay item i in base year 1987.

Q_i= total quantity of representative item i in period 1981-1997.

The authors also established submodels, depicted in Figure 3.3, to estimate the price of every pay item, which are: price of labor, price of material, price of equipment, pay item quantity, contract duration, contract location, quarter in which contract was let, annual bid volume, bid volume variance, number of plan changes, changes in standards or specifications.

Figure 3.3

Wilmot and Mei model for highway construction overall cost estimate



Note The diagram depicts the composition in submodels with pay items of the overall model employed by Wilmot and Mei for every pay item (Wilmot & Mei, 2005).

To represent each pay item, the authors chose to adopt a multi-layer feed-forward network whereas for training a backpropagation learning algorithm was used. The authors, after several changes to the network, reported that, overall, five neurons in the hidden layer constituted the optimal

number, with a tangent sigmoid transfer function[5] due to the binary inputs in the input dataset. Eighteen neurons were used in the input layer to represent the input variables, listed above, plus eight binary variables accounting for Louisiana's nine districts to represent the location factor. Lastly, in the output layer, only one neuron was inserted, to represent the unit price of the pay item modeled (Wilmot & Mei, 2005).

The model employed contract data from the Louisiana Department of Transport and Development from 1981 to 1997. Of this dataset, 85% of data was randomly selected for training, set to stop after 5,000 iterations or if an RMSE[6] of 0.01 was achieved, while the remaining 15% was employed in the testing phase. The overall model's testing phase was conducted with the data over the 1984-1997 period. Specifically, mean annual estimates of input variables from the 1981-1984 period, were employed in the submodels so that the final LHCI was estimated. From Figure 3.4 it is possible to notice the differences between the observed data and the estimation in the 95 percent confidence level. The authors applied the model to generate a forecast for the overall cost of Highway construction over the period 1997-2015, utilizing forecasts from the Bureau of Economic Analysis and the Directorate of Revenue Intelligence, along with average values observed over 1993-1997 for

---

[5] A tangent sigmoid transfer function is a neural transfer function, which calculates the output of a layer given its input. Operatively, it takes a matrix of net input vectors N, and returns the S-by-Q matrix, A, of the elements of N squashed into [-1;1] (*Hyperbolic Tangent Sigmoid Transfer Function*, n.d.). An S-by-Q matrix presents columns with only binary values [0,1]. The index of the largest element in the column indicates which of the S categories that vector represents (*Classification Confusion Matrix*, n.d.).

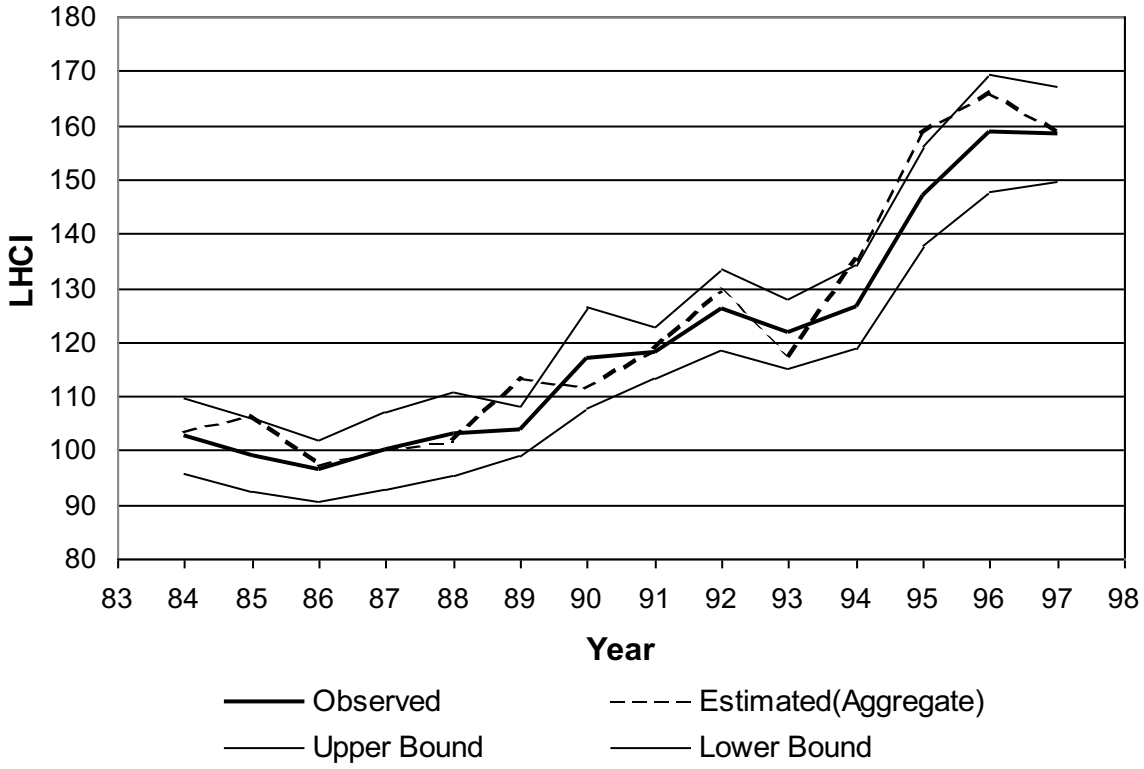
[6] $RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\check{y}_i - y_i)^2}{n}}$

The root mean square error is a measure of the average magnitude of error seen as the normalized distance between the predicted values and the observed ones. Where: i=index of the sample,
$\check{y}$=predicted outcome,
y=the actual value,
m=number of samples.

items such as quantity, contract location, and duration. From this estimate, a growth rate of approximately 3.4 percent per year was obtained, as shown in Figure 3.4.

Figure 3.4

Estimated and observed cost from Wilmot and Mei's study



Note From the graph, it is possible to notice that the estimate presents a 95% confidence level in the testing phase.

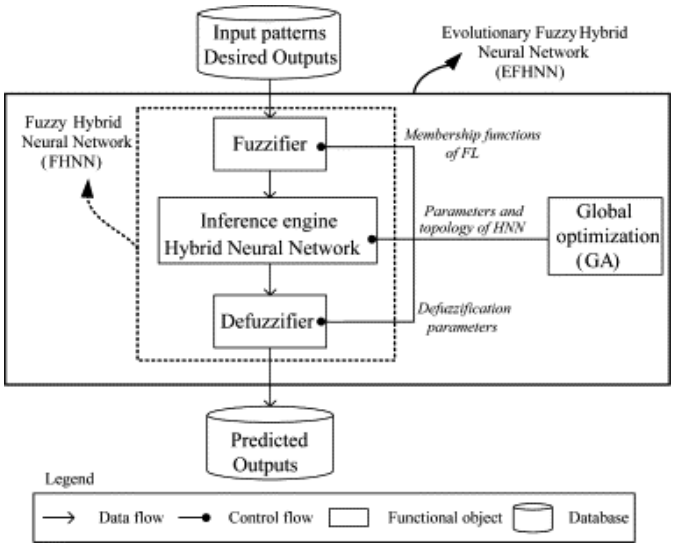
Fuzzy Neural Networks and Higher Order Neural Networks

As previously mentioned in a previous section, Fuzzy NN integrates ANN and Fuzzy Logic. One interesting example of this technique is the model presented by Cheng and Tsai in their work *'Conceptual cost estimates using evolutionary fuzzy hybrid neural network for projects in construction industry*'. They propose a combination of Fuzzy Logic (FL), Genetic Algorithm (GA), which will be analyzed deeply in a following chapter, and Higher Order Neural Networks (HONN).

The fundamental difference between ANNs and HONNs is that the latter perform a weighted multiplication of input variables instead of a weighted sum. They present increased adaptability and an easy display of how inputs are mapped into outputs, due to their specific structure and the use of a non-linear equation into a specified layer, which allows networks to capture higher-order correlations easily and obtain non-linear mapping effectively (Cheng et al., 2010). Higher-order neural networks may vary according to the type and order of operations performed, along with the presence or absence of the hidden layer and one changeable weight (Behera et al., 2023). Specifically, the authors developed the algorithm shown in Figure 3.5. Its general functioning sees each NN connection selecting either a linear or higher-order NN connector, and then the optimization through GA adaptation process, employed to search simultaneously for optimum FL membership functions, defuzzification coefficients, HNN topologies, and HNN parameters (Cheng et al., 2010).

Figure 3.5

EFHNN Architecture in Cheng et al.'s study



Note The figure clarifies the role of the diverse algorithms employed in the model, specifically the optimization role of GA and the fuzzy logic employed before and after the application of the HNN (Cheng et al., 2010).

The Hybrid Neural Network (HNN) used at the core of this study is a combination of a traditional neural network with a higher-order one, which for this specific case was the HONEST model, constructed of three layers with a high-order connection and a linear connection between the 1st and 2nd layer, along with 2nd and 3rd layer (Cheng et al., 2010). Additionally, the authors chose the apply high-order connections to all layers' connections. The following are the equations governing the neuron in the HNN where y_j is an HNN neuron output calculated by neuron inputs x_i. c_{ji} represents a coefficient of an interconnection, which can be in linear or high-order format based on the weight w_{ji} or exponent p_{ji}, respectively. An activation function f uses a sigmoid function with a slope coefficient of a. Therefore, each layer connection features an attached connection type that represents the corresponding operation selection.

Linear connection: $y_j = f(\ \sum wjixi\ +\ bj0\ \times 1\ )$

Higher Order connection: $y_j\ =\ f(\prod x^{p_{ij}}{}_i\quad \times\ 1^{bj0}\ )$

Activation function: $f(x) = \frac{1}{1+e^{-ax}}$

All HNN parameters are then optimized by GA evolution. As noted above, an HNN with 2 layers may select either a linear layer connection (*L*) or higher order connection (*HO*). Four possible scenarios, based on connection type, exist for 3-layer HNN models, including *L–L*, *L–HO*, *HO–L*, and *HO–HO*, leading to a total of $2^N - 2$ HNN model candidates, of which only $N - 1$ models select all *L* connections (Cheng et al., 2010). Regarding the Fuzzy Logic in the defuzzification layer, a membership function (MF) initially assigns inputs into one of several membership grades. In this study, a complete MF set using trapezoidal MF has been adopted.

The two estimating results planned by the authors were: an overall cost estimate, to be used in the planning stage in the absence of category estimates, and specific estimates for the latter. The authors identified six quantitative factors (Floors underground, Total floor area, Total floor area, Site area, Number of households, and Households in adjacent buildings) and four qualitative factors (Soil

condition, Seismic zone, Interior decoration, and Electro-mechanical infrastructure), all belonging to the planning stage of the project, as input factors for the overall cost estimation. For the seven category estimates for the engineering costs (Temporary construction, Geotechnical construction, Structural construction, Decorative construction, Electro-mechanical infrastructure, Miscellaneous construction, Indirect construction) the input factors can be consulted in Figure A.2 of the Appendix, to calculate the engineering cost. Overall, the algorithm displayed 1 input neuron, and 5 neurons in each of the five hidden layers, with a mutation rate of 2.5% and crossover rate of 90%. The authors employed 28 construction projects with data ranging from 1997 to 2001, with a US dollar per square meter cost ranging from 1242.1 USD/m^2 to 3038.4 USD/m^2, and allocated 23 cases for training and 5 for testing. The authors decided to compare the results of this algorithm with the results obtained with an evolutionary fuzzy neural inference model (EFNIM) developed by the authors in a previous study, which did not employ higher order neural network and changes to FL and GA (Cheng et al., 2009). As it is possible to notice from Figure 3.6, the EFHNN algorithm, while more complex, offered more accurate results, measured with RMSE[6]. Specifically, while the overall cost estimate error is still significantly lower than the traditional average for this category (Messner, 2019), other methodologies as already presented, offer better accuracy levels. It is also possible to notice that category estimates, given the increased number of input factors, sport a better accuracy (Cheng et al., 2010).

Figure 3.6

EFHNN and EFNIM application results' errors in Cheng et al.'s study

Case no.	EFHNN		EFNIM	
	Overall estimate error (%)	Total category estimate error (%)	Overall estimate error (%)	Total category estimate error (%)
1	19.312	0.900	20.541	2.504
2	13.187	5.452	23.783	7.458
3	5.797	4.349	21.201	9.699
4	3.336	11.447	5.082	10.018
5	10.148	7.373	9.755	4.082
Avg.	10.356	5.904	16.072	6.753

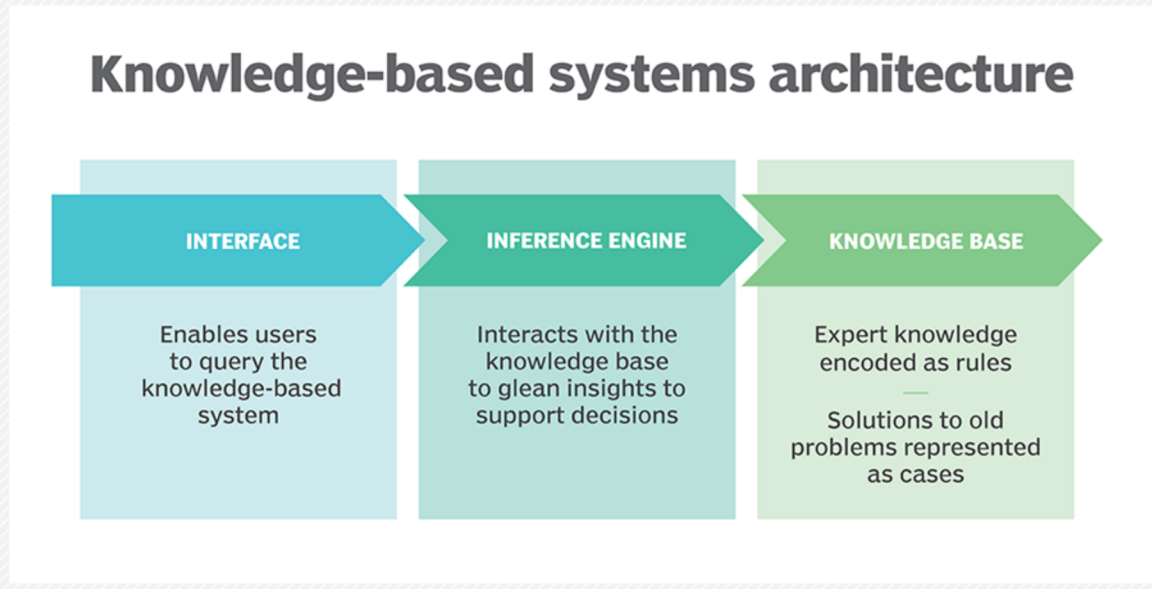
Note The table displays the accuracy of the different cases of Cheng et al. study with fuzzy neural networks for cost estimation (Cheng et al., 2010).

Knowledge-Based Systems

A Knowledge-Based System (KBS) in AI is a computer system that analyzes information from various sources to generate new knowledge (Gills & Moore, 2023). These systems possess problem-solving capabilities and can understand the context of the data they process, allowing them to make informed decisions based on stored knowledge. A typical Knowledge-Based System consists of three key components, as depicted in Figure 4.1: a knowledge base, an interface engine, and a user interface. The knowledge base serves as a repository for information and resources, while the interface engine processes data and retrieves relevant information based on user requests. The user interface enables interaction with the system, allowing users to submit queries and receive responses. Knowledge-based agents in AI operate at two levels: the knowledge base level and the implementation level. The inference engine processes and locates data in the knowledge base level, based on requests. Then, the reasoning system, which consists of the implementation level is used to draw conclusions from the data provided and make decisions based on if-then rules, logic programming, or constraint-handling rules (Gills & Moore, 2023). Knowledge-based agents utilize the existing knowledge and current inputs from the environment to infer hidden aspects of the current state before determining the appropriate action to take.

Figure 4.1

Knowledge-Based System Steps



Note The diagram shows the three main tiers of KBS architecture (Gills & Moore, 2023).

KBSs belong to Limited Memory AI, which makes informed and improved decisions by studying past data from its memory. However, they are not part of ML as they do not use mathematical models and pattern recognition to make their inferences but employ Boolean logic. This category includes several types of approaches, of which two of them can be employed for cost estimation: Expert Systems or Rule-Based Systems (ES or RBS), and Case-Based Systems (CBR).

Rule-based Systems or Expert Systems

Rule-based Systems or Expert Systems rely on human-specified rules to analyze or modify data to achieve a desired outcome. They simulate human expert decision-making in a specific field, offering solutions and explanations for problems. Expert systems are designed to solve complex problems by logic application through bodies of knowledge, employing IF-THEN rules in lieu of conventional procedural code (Sanni et al., 2022). They were the first algorithms employed in the early years of AI development. Indeed, Machine learning was not widely implemented until the late 90's when sufficient computational power was reached. Hence, for most of AI history, the algorithms

employed were rule-based or expert systems (Santosh et al., 2022). As already mentioned, the rules employed have an IF-THEN structure (condition and action), which are sent to the inference engine, that communicates with the knowledge base and employs a rule applier and a pattern matcher. The latter determines which learning rules are connected, aiding the former in the selection of the rules. The new information produced by the action becomes part of the working memory. This process is repeated to find relevant rules between the knowledge base and working memory. The main two rule systems employed are data-driven forward chaining, which infers based on initial facts, and goal-focused backward chaining, where the process starts on a hypothesis that needs to be proven. As the rules, being coded by humans, might contain uncertainty, various methods are usually employed to assign uncertainty values (Menaga & Saravanan, 2021). ES are typically employed when the knowledge base is available and not ambiguous or changing, but with a manageable size, and when the solution depends on logical reasoning and not intuition, which the system lacks.

Case-Based Systems

Case-Based Systems or Reasoning utilize past data from similar situations to develop solutions for a problem using case-based reasoning (Sanni et al., 2022). They will be further analyzed in the following section.

Knowledge-based systems for cost estimation in construction projects

Given the results of an extensive literature review, this thesis will focus on Case-Based Reasoning. Indeed, the literature analysis revealed the lack of recent studies on expert systems for cost estimation in construction projects, as they are mainly employed in hybrid models in combination with Machine Learning (ML) models for managing and estimating cost contingency, factors identification, and risk analysis.

The author suggests the idea that Case-Based Reasoning might be better suited for cost estimation in comparison to ES/RBS. They necessitate a well-known stable environment so that the theory or knowledge applied by the rules is consistent with it. This formalization of knowledge might

be extremely complex when the area of application is weakly theorized or volatile. Case-based reasoning systems, thanks to their capacity to leverage past experiences, avoid the formalization of knowledge in the form of rules. They also make it possible to call on the intuition, the judgment, and the habits of the expert, and to obtain a result or a decision, even when there is no theoretical model of the system in question. Furthermore, one of the advantages of the case-based reasoning technique is the facility of development, compared to rule-based systems (Duverlie & Castelain, 1999).

Case-Based Reasoning

Case-based reasoning (CBR) is defined as "an artificial-intelligence problem-solving technique that catalogs experience into "cases" and correlates the current problem to an experience" (Definition of Case-Based Reasoning (CBR), n.d.). CBR is rooted in the dynamic memory model of Roger Schank, dated 1982.

A CBR system presents this modus operandi (Zima, 2015):

-Creating a database containing costs of elements of previous construction projects, such as the ones listed in the first chapter of this paper, each described with diverse information like cost of work, the scope of works, date of realization, and location. The database structure needs to be hierarchical so that finding the desired construction elements is easier and the number of similarity comparisons is smaller.

-When a new case is presented, the CBR system retrieves one or more stored cases similar to the new case according to the percentage similarity (similarity score) calculated by a user-defined similarity function.

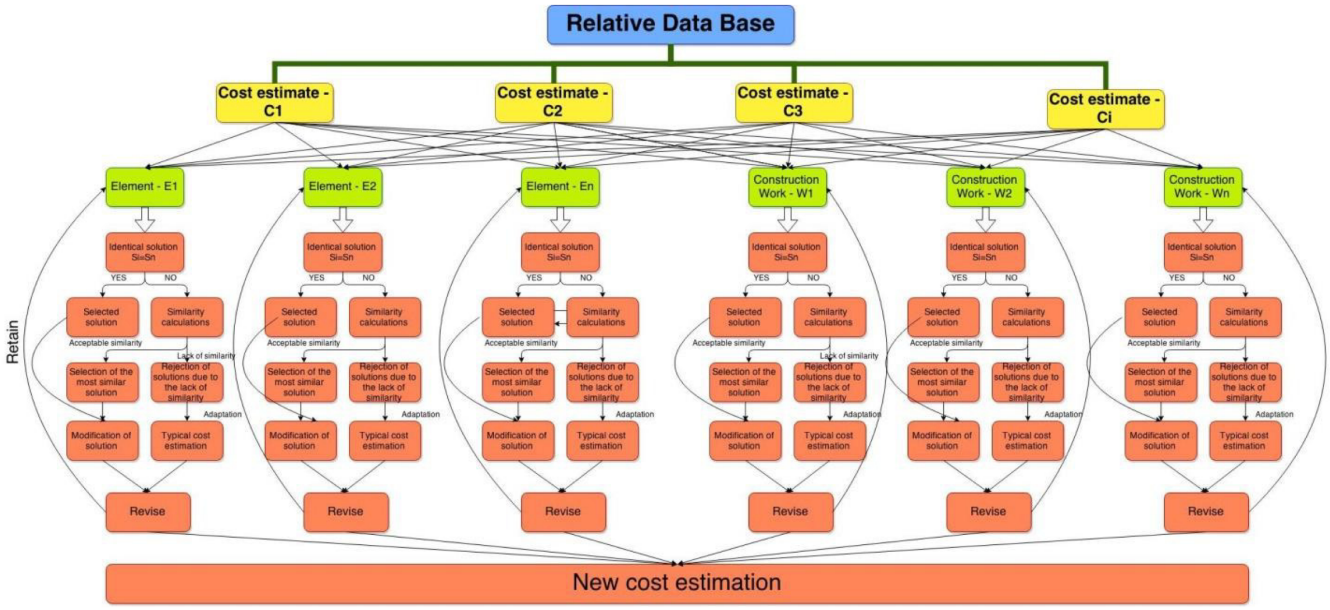
-Once similar cases have been retrieved there is the adaptation phase during which the solution is reached adapting the historical data to the new project to reach a tailored output. Models can implement automatic adaptation programmed by the creators of the model, or the user can conduct the adaptation.

-The new solution is retained as a part of the stored cases throughout the test.

Following the above description, which is exemplified by Figure 4.2, it may be said that CBR is similar to expert judgment and analogous techniques, which are based on comparison. However, with artificial intelligence, the analysis will be done by a machine, with an increased ability to manage larger volumes of information (Zima, 2015).

Figure 4.2

Cost estimation process with CBR in Zima's study



Note The diagram depicts the database structure employed in Zima's study (Zima, 2015).

There are different studies on CBR applied to cost estimation in construction projects, even if a significant amount applies hybrid models, with CBR combined with other techniques such as AHP in An et al. studies (An et al., 2007). Following a study conducted by Krzysztof Zima in 2015, a case study of the application of a CBR to the construction of a sports field in Alwernia, Poland, it might be possible to grasp the inner workings of this algorithm.

In the study, the database has been developed by breaking down the cost estimates presented by contractors in the tender procedure. Only selected offers have been included. Concerning more

information regarding the indexing procedure and method adopted by the author, please refer to the paper in question.

The first step for estimating a new case with the CBR model is to conduct the customary check of whether the database contains an identical case. If there is no such case, the similarities (SIM) between the new case scope elements and the main indexed groups in the database are calculated to identify the element stored in the database that is the closest considering the description parameters.

The similarity calculated for every element is a sum of products of weights and local similarities SIM material and SIM amount of works as exemplified by the following formulas.

$$SIM\left(V_N, V_j\right) = 1 - \frac{|V_N - V_j|}{V_{max} - V_{min}}$$

For the sub-criteria adopting numerical values within the range [V_{min} – V_{max}].

$$SIM\left(V_N, V_j\right) = 1 - \frac{n(V_N) - n(V_j)}{M - 1}$$

For the sub-criteria described by linguistic values, where:

V_N is the element from the new case under analysis.

V_j is the database element.

After this step, automatic modification of the selected solution is needed and, in this specific case, it required correcting the selected price by the indexing and regional coefficient. The final cost estimation for the new case is equal to the sum of construction elements unit cost estimated and

corrected, multiplied by the amount of work from the bill of work quantities. The global mean absolute estimation error (MAEE[7]) was equal to 0.057.

Considering that the mean calculation error in Poland at that time, viewed as the difference between the calculation prepared by the investor and the calculation of the chosen offer, ranges between 0.3 and 0.4, the result can be considered very satisfactory (Zima, 2015).

Following the analysis of another study it is possible also to assess the performance of CBR compared to other algorithms, such as Artificial Neural Networks (ANN), belonging to the family of ML, and linear regression. A comparative study of the actual construction costs of 530 projects of residential buildings built between 1997 and 2000 in Seoul, Korea, presents the construction costs estimated with a regression model, an ANN model with 12, 25, and 1 neuron in the input, hidden, and output layers, and a case-based reasoning model with 40 test data. The results gave Mean Absolute Error Rates (MAERs[8]) of 6.95, 2.97, and 4.81, respectively, as shown in Figure 4.3.

---

[7] Global Mean Absolute Estimation Error (MAEE) is a linear evaluation of error, obtained by calculating the absolute average distance between paired observations expressing the same phenomenon (Zima, 2015).

$$MAEE = \frac{1}{N}\frac{\sum_{i=1}^{n} | C_{CBR} - C_{ACT} |}{C_{ACT}}$$

Where N= total number of observations,
C_{CBR} = estimation of x_i,
C_{ACT}= observed real value.

[8]Mean Absolute Error Rate, which describes the absolute error rate (AER) and the frequency of AER within a specific range (Kim et al., 2004).

$$MAER = \frac{\left(\Sigma \left| \frac{C_e - C_a}{C_a} \text{ x } 100 \right| \right)}{n}$$

Where C_e=estimated construction costs,
C_a=collected actual construction costs,
n=number of test data.

Figure 4.3

MAER[8] results for the multiple regression, neural network, and case-based reasoning models implemented in the paper "Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning" (Kim et al., 2004).

Error rate (%)	MRA		NNs										CBR	
			Best model		25 models with 12-9-1		25 models with 12-12-1		25 models with 12-25-1		Average of 75 models			
	Fre.	Cum.	Fre.	Cum.	Fre.	Cum.	Fre.	Cum.	Fre.	Cum.	Fre.	Cum.	Fre.	Cum.
0–2.5	28	28	48	48	15	15	13	13	13	13	8	8	43	43
2.5–5	15	43	32	80	45	60	35	48	37	50	40	48	20	63
5–7.5	15	58	18	98	12	72	30	78	38	88	34	82	10	73
7.5–10	15	73	0	98	18	90	10	88	3	90	8	90	10	83
10–12.5	15	88	2	100	5	95	5	93	3	93	3	93	7	90
12.5–15	5	93	—	—	0	95	3	95	2	95	2	95	8	98
15–17.5	5	98	—	—	5	100	2	100	5	100	5	100	0	98
17.5–20	0	98	—	—	—	—	—	—	—	—	—	—	2	100
⩾ 20	2	100	—	—	—	—	—	—	—	—	—	—	—	—
MAER	6.95		2.97		5.61		5.80		5.54		5.65		4.81	

Note From the table it is possible to observe that the most accurate model was the ANN with 12 input nodes, 9 hidden nodes, and 1 output node, while the multiple regression model reported the highest error of all (Kim et al., 2004).

From these results, it can be asserted that the ANN model is the most accurate one, but it requires a long trial-and-error process. Taking into consideration the practical implementation of these models as well, the tradeoffs between time and accuracy ought to be evaluated. Concerning this aspect, the most viable approach is CBR as it requires less time to devise and, also, has a superior quality of explanation of the estimation process, as ANN models offer only black-box solutions (Kim et al., 2004). Specifically, the similarity was calculated by the authors with the following formula:

$$\text{Percentage Similarity}(\text{N}, \text{S}) = \sum_{i=1}^{n} f(N_i, S_i) \times w_i \sum_{i=1}^{n} w_i \times 100(\%)$$

Where N is the new case, S is the historical case in the database, n is the number of variables in each case, f is the similarity function for variable I in cases N and S, and w_i is the importance weight of variable i.

The similarity function was determined with the following formula:

$$N_i - S_i S_i \times 100 \leq 10(\%)$$

where N_i was the value of the new case's variable, S_i was the value stored case's variable, and 10%, was the matching range, empirically determined by the authors, that describes the matching tolerance between the values of the variables of the new and stored.

If the value of the new case's variable respected this equation, the value of $f(N_i, S_i)$ was 1, otherwise, it was 0. For the weight (w) of variables the gradient descent method[9], an optimization algorithm, was chosen. After the percentage similarity of all the cases had been calculated, the cost data in the case base were ranked. In this study, the top-ranked case was selected and the cost corresponding to this case was the final estimation for the construction project.

[9] Gradient descent is an optimization algorithm based on the minimization of the loss function of the prediction. It relies on two components: direction, determined by the partial derivatives of the loss function, and the learning rate, which sets the size of the steps taken towards reaching the minimum (What Is Gradient Descent?, n.d.).

Evolutionary Systems

Evolutionary algorithms are part of the evolutionary computation approaches, in which "solutions, instead of being constructed from first principles, are instead evolved through processes modeled after the elements of Darwinian evolution" (Altenberg, 2016). These techniques are a heuristic-based approach to problems that cannot be easily solved in polynomial time due to a significant number of variables.

The general functioning of these methods sees the trial phase as the generation of candidate solutions and the evaluation of the error between them and the expected outcome. The error is then used to identify which solutions should generate a new batch of results. Evolutionary computing encases methods such as evolutionary algorithms and programming, and swarm intelligence models (*What Is Evolutionary Computation?*, n.d.).

Operatively, the evolutionary algorithm employs a population of solutions whose effectiveness is evaluated through a fitness function. The ecosystem evolves over time, generating fitter and fitter members, routinely reaching a more accurate outcome (Leach, 2006). In fact, with this algorithm, fitter members will determine the next generation, cutting off solutions deemed not useful from determining the genetic pool of the future generations (Soni, 2018). Analyzing the mechanism more deeply, it is possible to comprehend the different steps, shown in Figure 5.1:

- Initialization, where the initial population of possible solutions, called members, is created, often randomly. The population must encompass a wide variety of components as it constitutes the gene pool of the algorithm, through which different possibilities will be explored.

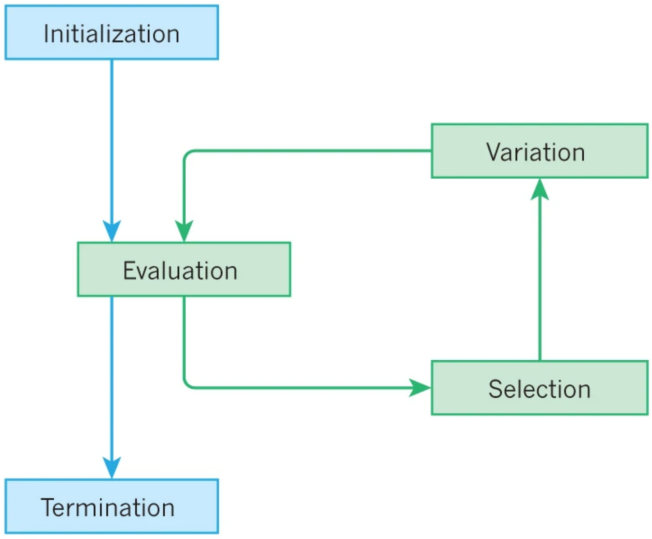
-Selection, in which members are subjected to an evaluation through a fitness function, that takes in the characteristics of a member, and outputs a numerical representation of how viable of a solution it is. Only a portion of the most viable solutions will be selected.

-Genetic operators, where the solutions selected are combined (Crossover) to create a new generation of solutions, applying variation operators to change the new members (Mutation) with a certain percentage of probability.

-Termination, when the algorithm stops either because the maximum runtime or maximum threshold of performance has been reached (Altenberg, 2016).

Figure 5.1

Evolutionary Algorithms' processes



Note The figure displays the main steps and phases of a generic evolutionary algorithm, highlighting the selection, and variation (Eiben & Smith, 2015).

Evolutionary algorithms are grouped into different classes. The main ones are Genetic Algorithms (GA); Evolution Strategies (ES); Differential Evolution (DE) and Estimation of Distribution Algorithms (EDA).

Genetic Algorithms

Genetic Algorithms are widely popular amongst evolutionary algorithms. Their solutions, often referred to as chromosomes, are vector items of decision variables, which tend to represent a value that will undergo optimization. GA solutions are usually represented as fixed-length vectors.

The mutation factor usually ranges around 4-8% and works by taking any of a discrete set of k possible values of the k-sized array and choosing a random new value from the available alphabet to replace it with. In the case of real-valued parameters, the replacement value is taken from a uniform distribution in a range, or a non-uniform probability distribution centered around the current value. Recombination, on the other hand, is typically implemented using single, two-point, or uniform crossover. Single-point crossover selects a random point of two parent solutions, swapping the remaining halves, to obtain two new elements (Kaya et al., 2011). Two-point selects instead two parents and two crossover points within them. The new solutions are created by swapping the values of decision variables between the two crossover points. Uniform crossover, conversely, follows a similar path except for crossover points, which are created following a given probability(Corne & Lones, 2018).

Regarding the selection of the individual for the mating pool, rank-based; tournament method; or roulette selection are typically employed. Rank-based selection involves ranking the population in terms of their fitness function, giving each member the chance of becoming a parent proportional to their fitness value. This operation can be very time-consuming but can prevent premature convergence (Basak, 2018). In tournament selection, conversely, a small group of solutions are uniformly sampled from the population, and their fitness function is compared. The winner of this tournament is the individual with the highest ranking. By adopting this method, the mating pool, consisting of only winners of these tournaments, displays a higher average fitness than the average population fitness, which leads to an increased average fitness function for the next generation(Ille R ’ & Goldberg, 1995). Lastly, roulette selection sees the fitness values of every member normalized and distributed. The group of values, representing new probabilities are then inserted in a pool, that is usually represented as a spinning wheel. A random spin of the wheel determines the selected individual for reproduction, according to their fitness rank, which now represents its incidence probability (*What Is Roulette Wheel Selection in Genetic Algorithms?*, n.d.).

Evolution Strategies

Evolution Strategies share significant similarities with Genetic Algorithms. Nevertheless, whereas the latter mutate only the mating pool factors to generate a new population, evolution strategies mutate every decision variable at the application of the mutation operator, determined by a mutation strategy, which oftentimes determines the probability of generating new individuals. Usually, strategy parameters are adapted during the application of the algorithm, as different mutation types are beneficial at different search stages. Currently, several methods are employed, from deterministic rules to self-adaptation ones. The most common one is the covariance matrix adaptation, which estimates productive gradient directions in the search space and adjusts mutations accordingly.

Another key difference between evolution strategies and genetic algorithms is the recombination operators, the use of multiple parents to create each child solution, and deterministic selection mechanisms, always employing the best solutions for the creation of the future generation (Corne & Lones, 2018).

Differential Evolution

Differential Evolution is contrastingly unique in the family of evolutionary computation, as it does not use probability distributions. Like Evolutionary Strategies, it differs from GA for their mutation operators, as it employs a geometric approach, which involves "selecting two existing search points from the population, taking their vector difference, scaling this by a constant F, and then adding this to a third search point, again sampled randomly from the population"(Corne & Lones, 2018). The mutated vector is then recombined with a target vector and replaced by the child solution obtained if its value is greater, an action that acts as the selection mechanism. A distinct peculiarity of this algorithm is it then that the population is slowly replaced with the new population. This algorithm is mostly self-adapting, and it involves fewer parameters leading to an easier implementation than, for example, covariance matrix adaptation.

Estimation of Distribution Algorithms

Lastly, Estimation of Distribution Algorithms implement the use of probability distributions, employing them not to describe the distribution of the next moves, but the composition of the next generation. They follow the traditional structure of EA, but after the selection process, a distribution that models the high-ranking solutions of the population is constructed, and the new solutions are generated by sampling the new distribution created. The distributions employed are usually chosen to satisfy a certain trade-off between efficiency and expressiveness, with the spectrum going from Bayesian networks, which can be expensive to construct by that capture dependencies well, to univariate distributions, easy to build but unfit to identify dependencies (Corne & Lones, 2018).

Evolutionary Systems for Cost Estimation in Construction Projects

In cost estimates, genetic algorithms are usually employed, often in combination with others to optimize the solutions, as already mentioned in Chapter 3. It is of interest a paper from Kunming University of Science and Technology, in which a Least Squares Support Vector Machine (LSSVM) is combined with a GA. LSSVM is a machine learning method based on the principle of structural minimization between points of the same class (Samui & Kothari, 2011). This specific model presents, compared to traditional SVM, the function to be minimized as the sum of squared errors. This algorithm was chosen given its ability to manage multivariable systems without the need to define a distribution, which suits well the task of project cost estimation. The choice to employ a GA was due to the need for optimization and diversity (Xu et al., 2015). The overall aim of the study was to generate a forecasting model for cost estimates for residential construction projects using cost information from Jiangsu Province, China, published from 2014 to 2015. Specifically, they utilized the genetic algorithm for feature selection, removing irrelevant or redundant features, improving the accuracy of the model, and cutting the solution time by evaluating feature subsets, combinations of multiple features, and their impact on the prediction of the target variable. Indeed, the random item

of the population was a M-array, composed of binary features (1 selected, 0 not selected), with M as the number of features. The fitness formula employed is the following:

$$f(x_i) = \frac{1}{\lambda_1 * MAPE\ (i)\ + \lambda_2 * \frac{n * feature(i)}{M}}$$

Where M is the number of test samples, λ_1 , λ_2 are user-defined parameters, which are instrumental in balancing the prediction accuracy and the number of features of the fitness function. The overall aim was to minimize the mean absolute percent error (MAPE[10]), as to achieve better accuracy of the estimation. Regarding crossover and mutation, the researchers employed the roulette method with a single-point crossover and a cross-point position randomly selected (Xu et al., 2015).

Over a total of 42 studies, 38 were employed for the training phase and 4 for the test phase. The input data selected to generate the output of cost per square meter are the construction area, foundation type, foundation bottom elevation, structure type, seismic intensity, roof waterproof, roof insulation, façade, wall insulation, windows and doors, ceiling, number of stories, ground floor, wall insulation, number of floors, floor height, top of wall level, concrete price, and steel price. To avoid distortions given by significant dimensional differences, the input data was normalized and the correlation between each input and the output variable was computed, shown in Figure A.3 of the Appendix. The authors decided to select the best input variables by running three different experiments, running the LSSVM with different input feature sets. One with features with correlation higher than 0.15; one with features obtained by the application of the genetic algorithm with an initial population composed of all the features and with an optimized fitness function; the last one is

[10] MAPE is an error measure defined as $MAPE = \frac{\sum_{i=1}^{m} |y_i - \tilde{y}_i| / y_i}{m} 100$ where:

y_i =actual cost;
$\tilde{y}_i$= estimated cost;
m= number of test samples;

comprised of the population generated by the previous GA further refined by another GA with an initial population size of 20, 30 maximum runs, crossover equal to 0.6, and mutation probability 0.02, set of the number of high-ranking features extracted from the correlation sort list C as 30% of total features. The resulting MAPEs[10] for the three trials are displayed in Figure 5.2. As it is possible to notice, the third trial presents the lowest resulting MAPE[10] of 6.47% (Xu et al., 2015).

Figure 5.2

Table of the resulting errors for the three trials of Xu et al.'s hybrid model experiment

Feature set	MSE	MAPE
F1	17415.52	7.76%
F2	14750.10	6.94%
F3	12097.81	6.47%

Note The table lists the accuracy results of the three trials with different input variables, clearly showing the superiority of the third trial, employing GA-selected features (Xu et al., 2015).

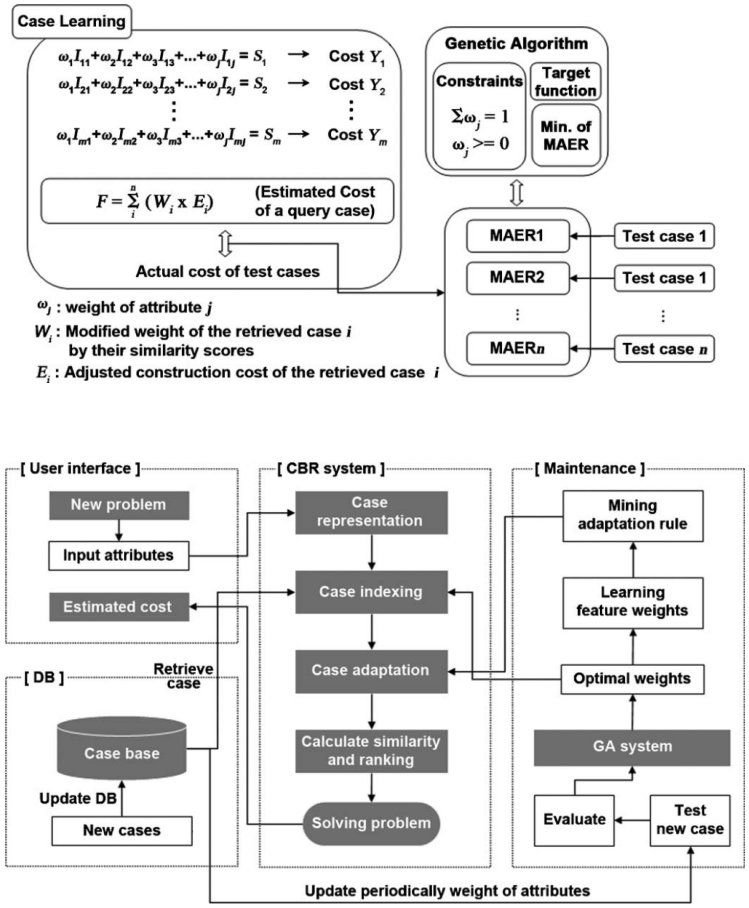
Another interesting study shows the application of CBR with GA, specifically to determine the attributes' weight for the selection of similar cases. This issue has been addressed also with other methods, such as the already mentioned AHP, but this method is still based on expert experience, requiring human interaction, and thus uncertainty. The study *'Preliminary Cost Estimation Model Using Case-Based Reasoning and Genetic Algorithms'* (Kim & Kim, 2010) aims to develop an estimating method for early-stage planning of bridge construction, which is traditionally done with cost per unit area according to the type of bridge, that requires the evaluation of different alternatives. As such, during these early stages, the information available is very limited with the length and width of the bridge and the number of lanes, and more information is added only with the progression of the project (Kim & Kim, 2010).

The study data comprised the complete group of national road construction projects completed from 2000 to 2005 in Korea (Kim & Kim, 2010). The algorithm developed in the study, whose process

and architecture are shown in Figure 5.3, sees the application of the GA for the impact factors' weight estimation to overcome the limitation of using the conventional linear planning method, given the nonlinear relationship between the error rate and the factors' similarity scores (Kim & Kim, 2010).

Figure 5.3

The model developed in Kim et al. study



Note From the diagram, it is possible to notice the contribution of the GA and CBR, along with the output results (Kim & Kim, 2010).

Specifically, the factors' weights are employed to calculate the cases' similarity scores by multiplying their factors' similarity scores with their weights following this formula:

$$S_i = \sum_j^n (I_{ij} \times w_j)$$

Where i = case identification number; j = attribute identification number; n= the number of cases, S_i= similarity value of case i; I_{ij} = similarity value of attribute j; and w_j= weight of attribute j.

The objective function devised is, as seen in similar approaches, minimizing the MAER[8] previously mentioned. The algorithm functioning follows the already analyzed CBR, with cases described by their attributes or factors that impact the case costs (Type of bridge, Local area, Length, Number of lanes, Width, Area of slab, Position (on land/over river), Number of spans, Length of span, Length of pier, Type of foundation (base)). The similarity scores calculated during the application of the CBR, if text described are either discrete (0 or 1000), and if numerically described follow the percentage error between the new case and the stored one.

Regarding the Adaptation phase, the costs of similar cases are adjusted by taking into consideration the attributes of the case under analysis and according to the length, width, and construction cost index of the bridge. The following results are then multiplied by the similarity weight of the retrieved similar case W_i:

$$W_i = \frac{S_i}{\sum_{i=1}^n S_i}$$

W_i = similarity weight of the retrieved similar case i; and n = number of retrieved similar cases.

The weighted average of the construction cost from the retrieved cases is then used as the estimated construction cost for the newly learned case, and the MAER[8] is calculated.

Concerning the application of the genetic algorithm, the researchers decided to employ the Evolutionary Solver available on Excel. They employed a population that was gradually increased from 50 to 250, with reported efficiency reached at 200. The mutation and crossover rate were

increased from 0.075 to 0.5. GA showed its efficiency and its best solution when the mutation ratio was 0.25.

Following the application of the GA and their optimal weights, shown in Figure 5.4, it was discovered that the previously mentioned attribute called 'Positions (land/water)' is irrelevant in the estimation accuracy at the beginning of the project, when few attributes are known, and their distribution changes when more features are available.

Figure 5.4

Weights' values in Kim et al.'s study

Attribute	Number of lane	On land or over water	Slab area	Length	Width
Weight	0.290171	0	0	0.709829	0

Attribute	Number of lane	On land or over water	Slab area	Length	Width	Number of span	Length of span	Length of pier	Owner (local area)	Type of foundation
Weight	0.264696	0.024774	0.058408	0.1025481	0.089571	0.027473	0.033582	0.28671106	0	0.112237

Note The table lists the weights of the factors analyzed through Evolutionary Solver and reveals the changes in importance with the increase of attributes' availability (Kim & Kim, 2010).

To further test the results, the researchers decided to traditionally estimate the costs with the unit price per length and type of bridge and compared it, calculating the $MAER^8$, with the estimation obtained with the algorithm. The values, shown in Figure 5.5, reveal the superior accuracy of the presented algorithm, with a $MAER^8$ of 7.621% compared to the almost doubled ones of traditional methods. This result was obtained when the selection process selected "similar cases up to the third most similar with 15% similarity evaluation criteria" (Kim & Kim, 2010).

Figure 5.5

Error rates between traditional estimation methods and the presented algorithm in Kim et al.'s study

	Error rates		
Error rate	CBR (%)	Guide for preliminary feasibility study (by $/m^2) (%)	Guide for preliminary feasibility study (by $/m) (%)
Mean absolute error rate	7.621	12.26	15.76
Maximum error rate	16.419	41.16	63.68
Minimum error rate	0.314	0.36	2.63

Note As it is possible to notice from the table, the errors reported for the employed algorithm are significantly lower than the traditional ones (Kim & Kim, 2010).

Results and Discussion

Results

The presented studies and topics might not appear as a systematic review, given the ample space dedicated to general descriptions of the numerous algorithms. Indeed, as stated in the Introduction and Methodology section, the overall aim of this paper was to provide an overview of AI algorithms for cost estimation in construction projects for project managers, who may not be familiar with AI. As it is possible to notice now, the methodology adopted followed a systematic review selection process, but then only a few selected articles were analyzed in this paper with the ultimate objective of illustrating the inner workings of the analyzed algorithms to a reader without prior extensive knowledge of the AI field. The comprehensive lists of studies can be found in the Appendix, under Table A.4.

Part of the employed studies have been included in previous systematic reviews which can be found in the literature. Specifically, the review *'Cost Estimation and prediction in construction projects: a systematic review on Machine Learning Technique'* (Hashemi et al., 2020) mentioned the studies by Wilmot and Mei for ANN (Wilmot & Mei, 2005), by Cheng et al. for Fuzzy NN from 2010 (Cheng et al., 2010), by An et al. study from 2007 (An et al., 2007), by Kim G. et al. for the comparison of ANN, CBR and RA from 2004 (Kim et al., 2004), and by Kim and Kim for the highway projects for the analysis of GA from 2010 (Kim & Kim, 2010). Kim et al.'s study along with An et al.'s have also been included in the review focused on CBR by Hu et al. from 2016 (Hu et al., 2016), and in Castro Miranda et al.'s review from 2022 (Castro Miranda et al., 2022) along with Cheng et al.'s study from 2010. The author then illustrated Zima's construction study with CBR (Zima, 2015), and Xu et al.'s study for residential construction projects' cost estimation with LSSVM and GA (Xu et al., 2015). The author also included part of studies such as Mohamed and Moselhi's study from 2022, (Mohamed & Moselhi, 2022), Atapattu et al.'s review (Atapattu et al., 2023), Doğan et al.'s study from 2006 (Doğan et al., 2006), Ji et al.'s study from 2010 (Ji et al., 2010), Cheng et al.

from 2009 (Cheng et al., 2009), Koo et al.'s study from 2010 (Koo et al., 2010), Duverlie and Castelain's study from 1999 (Duverlie & Castelain, 1999), and Juszczyk et al.'s study from 2018 (Juszczyk et al., 2018) for clarification purposes on certain topics. Of these, only the studies from Duverlie and Castelain, Juszczyk et al., and Mohamed and Moselhi have not been included in Hashemi et al. and Hu et al.'s reviews.

In addition to these studies the author employed two other previous systematic reviews, the already mentioned one from Hashemi et al., Hu et al., and Castro Miranda et al.'s.

The articles non directly illustrated in the paper were excluded, as it is possible to notice from Figure A.1, due to their complexity or specific focus on optimization, as the main objective was clarity and readability of the content for non-experts of the AI field.

The selected studies and reviews did not report a risk assessment. Hashemi et al. reported the limitation coming from exploring the articles published in English only from two databases: Google Scholar and ScienceDirect, thus not considering all the articles outside of these groups (Hashemi et al., 2020). Thus, the author can only identify the same biases that have been identified for this thesis, mentioned in the Methodology. They are informational bias, observer bias, and cognitive biases such as confirmation bias and availability heuristics. However, it is also important to remember the threat of sampling bias, which could have affected the data employed in all the studies mentioned if the researchers have employed sampling data either too small or not representative of the population under analysis (Géron, 2019).

The following Table 6.1 displays a summary of the studies analyzed and their results.

Table 6.1

Summary of the accuracy results of the studies analyzed in the previous chapters

Authors	**Name**	**Algorithm**	**Objective**	**Confidence level**
Wimot, Mei	'*Neural network modeling of highway construction costs'*	ANN	Estimation of the overall cost of Highway construction.	RMSE[6] 0.01
Cheng, Tsai, Sudjono	*'Conceptual cost estimates using evolutionary fuzzy hybrid neural network for projects in construction industry'*	Hybrid EFHNN with HN, FL, and GA	Estimation of overall costs (pre-design stage) and category costs (design stage) for a residential building.	RMSE[6] 0.059
Zima	*'The Case-based Reasoning Model of Cost Estimation at the Preliminary Stage of a Construction Project'*	CBR	Estimation of construction costs of a Sports Field.	MAEE[7] 0.057

Kim G. et al.	*"Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning"*	CBR	Estimation of construction costs for a residential building.	MAER[8] 0.048
Xu et al.	*'Construction Project Cost Prediction Based on Genetic Algorithm and Least Squares Support Vector Machine'*	Hybrid LSSVM with SVM and GA	Estimation of construction cost for a residential building.	MAPE[10] 0.065
Kim and Kim	*'Preliminary Cost Estimation Model Using Case-Based Reasoning and Genetic Algorithms'*	CBR with GA	Estimation of construction cost of a bridge	MAER[8] 0.076

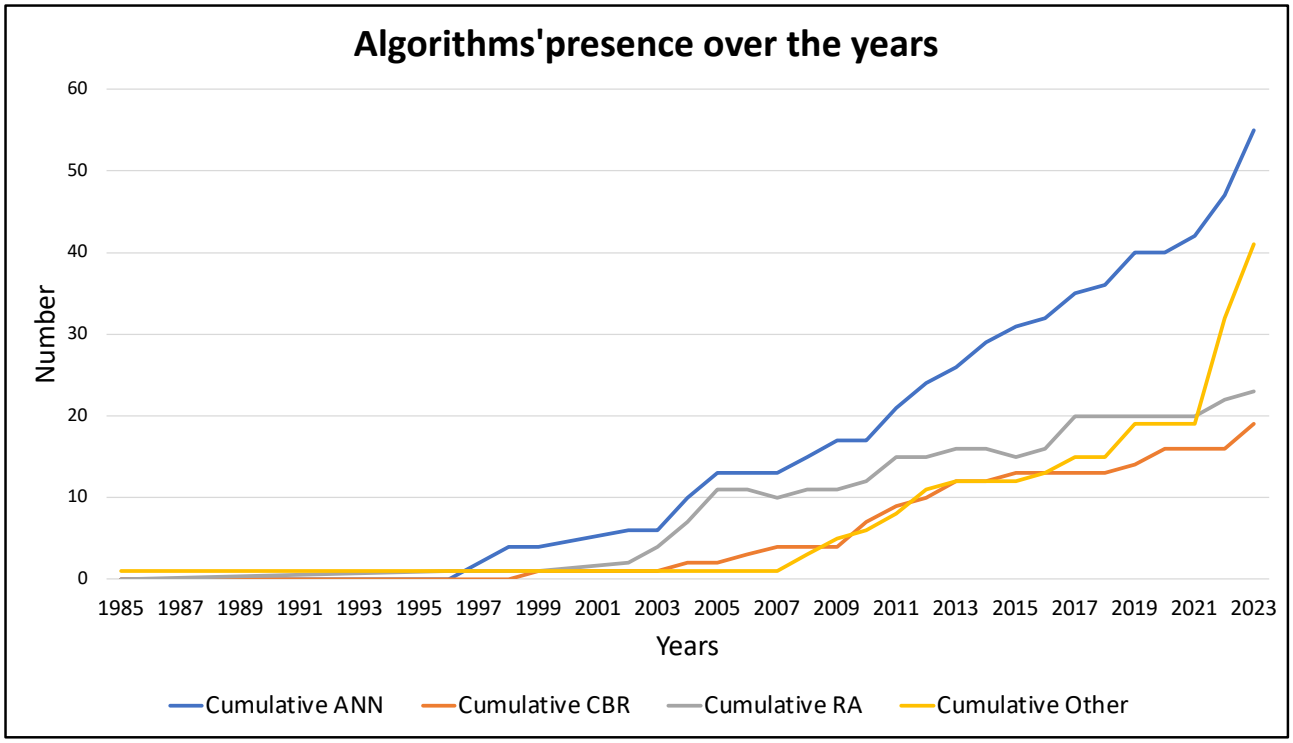
Discussion

Machine Learning

The studies presented and the literature review revealed that this category is the most studied. Indeed, Hashemi et al. reported in 2020 that machine learning algorithms where the most researched, specifically ANNs (Hashemi et al., 2020). This trend continues today, as it is possible to notice from Figure 6.1.

Figure 6.1

Methods' distribution in the literature review



Note The graph represents the distribution of the different methods employed in cost estimation studies for construction projects.

More specifically, as already presented in Figure 3.2, ANNs have been, since 2007, the more popular algorithms. In fact, over the last few years, their increase saw an exponential growth, with the use of optimization algorithms for weights and feature selection, which now constitute the new default approach in literature studies. It is also of interest also noticing the recent exponential increase

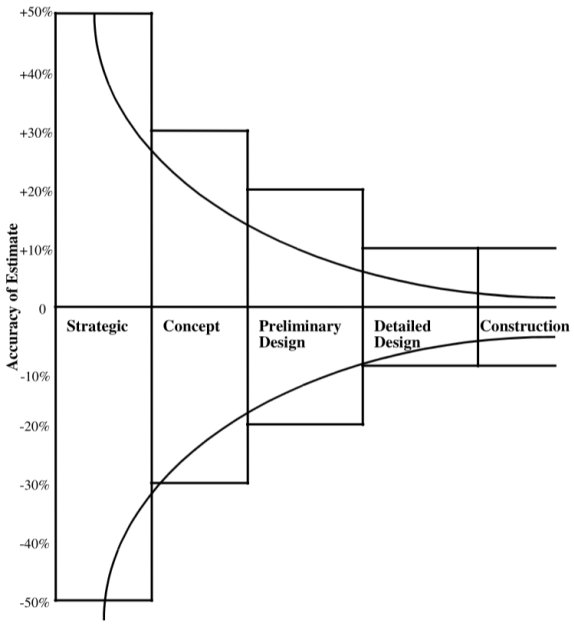
in the implementation of other models, mostly other machine learning models such as SVM, Gradient Boosting, and RF, that nevertheless are more complex to train and model (Géron, 2019).

The overall increased accuracy brought by the use of ANNs already in the strategic phase of a project, can be compared to the Detailed Design and even Construction phase, shown in Figure 6.2, as the two presented studies' errors report (Wilmot & Mei, 2005)(Cheng et al., 2010). In addition, both studies attest to the superiority of this approach's accuracy compared to traditional methods or RA, which, as Figure 3.2 and 6.1 attested, are still analyzed in research. Their employment saw a swift increase over the years 2003-2005 and proceed to today with a lower slope.

Kim et al.'s study, clearly demonstrates the increased accuracy of ANN methods, compared to CBRs and RA, with a final accuracy of 0.029 against 0.048 and 0.069 respectively.

Figure 6.2

Accuracy of cost estimates at different stages of the design



Note The diagram depicts the evolution of construction projects' accuracy (Abourizk et al., 2002).

Overall, ANNs increased accuracy is due to their ability to identify complex relationships among various project parameters (Cheng et al., 2010), along with their adaptability to diverse

construction projects and data sets. ANNs rely on data-driven learning, allowing them to adjust to evolving datasets and changing conditions, which contributes to improved predictions over time. Additionally, ANNs are effective at capturing high-order correlations between variables, further enhancing their predictive power.

Conversely, ANNs are exceedingly sensitive to input data. On one hand, a larger amount of data can increase the number of relations discovered and modeled. However, with the increase of input factors, the complexity increases, impinging on the accuracy of the results. Moreover, the number of neurons in the hidden layers along with the variable weights are decisive too. This is why, hybrids models that employ fuzzy logic and GA for parameter and weight selection are becoming more popular, such as the presented EFHNN by Cheng et al. (Cheng et al., 2010). The black box nature of ANNs makes interpreting and justifying their decisions difficult and can be limiting, particularly in cases that require transparency (Hashemi et al., 2020). Additionally, their substantial computational demands and training complexity may restrict their use in projects with limited computing capabilities (Kim et al., 2004).

Knowledge-Based Systems

The research process revealed a significant amount of papers focused on the application of KBS algorithms to construction projects for cost estimation, along with several studies investigating the application of blended algorithms or other areas of construction projects, such as contingency estimation, risk assessment, and management, findings backed up by review conducted by Hu et al. in 2016 (Hu et al., 2016). In the author's opinion, this attests to the interest that the scientific community has been displaying in this category of AI for cost estimation, even if the recent trend seems to be focused on ANN, as shown in Figure 6.1. Nevertheless, it is evident that the rate of application of CBRs underwent a significant rise since 2009 and to this day still sports an increase rate higher than RA, meaning a major interest compared to the one for the latter, justified by their superior performances.

The two studies presented and analyzed convey the superior accuracy of the estimates obtained with KBS compared to traditional techniques, and the ease of use of these methods compared to other AI approaches, such as ANNs. Both studies report an accuracy level of approximately 5-6%, comparable to a unit price estimate (Messner, 2019), as it is also possible to see from the diagram in Figure 6.2. However, as it is possible to infer from the Figure, the traditional accuracy for construction projects might reach an error of 5% only in the detailed design or construction phase, especially regarding underestimation. As ANNs, the advantage of Case Base Reasoning is that it can be used as soon as the strategic phase estimates, but it is significantly less time-consuming than ANNs or traditional approaches (Hashemi et al., 2020), especially if the algorithm has already been previously employed, as elements as the knowledge base, and the automatic adaptation program are already compiled. Traditional Estimates require several days to complete, even with a database of similar cases available. Moreover, in their comparison study, Kim et al. demonstrated that CBR is less time consuming than ANNs. Indeed, by comparison, ANNs require significant modeling, training, and updating time. It is true that to be effective, CBR needs a database with a significant amount of data in the form of previous bids. However, while ANNs needs to be remodeled and retrained, CBR is updated by feeding the new case to the system, which will calculate a similarity indexes, naturally lower, and store the case for future estimates (Kim et al., 2004). Additionally, according to a comparative study between parametric and CBR estimates, CBR guarantees transparency, which cannot be obtained with parametric approaches or black box algorithms. With CBR, the user can thoroughly investigate the solution process and possibly apply corrections (Duverlie and Castelain,1999).

Additionally, the knowledge base functions as the collective memory of the enterprise, storing all the information regarding previous bids and estimates generated and thus transforming knowledge that has traditionally been implicit, in the minds of experts, into explicit. Thus, they allow estimators

to employ the superior computational and memory power of a computer to achieve better estimates by comparing a significantly increased number of cases and previous estimates.

Moreover, the research revealed the widespread implementation of optimization strategies when using CBR, to optimize the algorithm, either in the retrieval phase of indexes and weights determination or case revision. It is of particular interest to note that according to Hu et al.'s review, research is more focused on weight optimization, given the fact that for index selection the human brain still outperforms automatization. Regarding the algorithms employed for weight selection, the most employed algorithms are GAs along with other possibilities such as RA, ANN, DT, feature counting (FC[11]), and gradient descent method (GDM[9]). Suggestions of more transparent methods have also been proposed, with the study of An et al. that employs AHP for optimization (An et al., 2007). Regarding their accuracy, Hu et al.'s review revealed that the most accurate ones are GA and ANN, even if they do not include a comparison study between the two. Conversely, Ji et al. highlight the fact that the performance of these trials varies greatly according to the model design and the combination of indexes (Ji et al., 2010), as such, as ANNs, CBRs retain a sensitivity that impinges on their accuracy.

Evolutionary Systems

Research on the application of evolutionary algorithms in the context of cost estimation for construction projects revealed an increased interest in their application, as stated before, for optimization purposes. As an example, in the case of CBR, genetic algorithms are frequently employed to optimize the index weights. Several studies have analyzed these methods' performance, with GA marked as the more accurate along with ANN (Hu et al., 2016). While there is a growing interest in the scientific community regarding the use of evolutionary algorithms, it is essential to

[11] In Doğan et al.'s and Koo et al.'s specific studies, Feature Counting consists of attributing a unit weight to all the variables selected (Doğan et al., 2006) (Koo et al., 2010).

acknowledge the relatively limited number of studies compared to other methods like Artificial Neural Networks (ANN). This limitation might present a potential bias, necessitating further analysis over time and in active project scenarios to ascertain the broader applicability of these algorithms, especially because the model structure, weights, and index deeply impact its accuracy.

Given the optimization purpose of evolutionary algorithms, it is of interest to report their accuracy compared to other optimization approaches. An illuminating study from Doğan et al. focuses on the comparison of GA, feature counting, and GDM for CBR's attributes weights optimizations. The study's output was the cost of the structural system for a residential building in Turkey, with all the variables affecting the structural system defined in the design stage employed as input attributes (Doğan et al., 2006). The report reveals that of the three, the most accurate was indeed GA, with 0.162 of average error, as possible to notice in Figure A.5 in the Appendix. Several other studies, reported in Hu et al.'s review, proceed with the comparison of different optimization methods for weights' definition in CBRs, and as already reported, while a direct comparison between GA and ANN-based optimization is not reported, these are the two most performing algorithms (Hu et al., 2016). The study '*A CBR-based hybrid model for predicting a construction duration and cost based on project characteristics in multi-family housing projects*' (Koo et al., 2010) highlights the difference in employing a feature counting-based optimization and a GA-based one, with even differences of 10 percentage points in accuracy between the two approaches, as Figure A.6 in the Appendix displays.

Overall, it can be summarized that genetic algorithms are indeed instrumental in the optimization phase of other algorithms, especially CBR. Nevertheless, they are also applied to other models, such as ANN-based ones, as Cheng et al.'s study reveals (Cheng et al., 2010). The fact that the literature is rich in studies employing GA for weight optimization in CBR-based estimates, can be logically explained by the inferior performance of this method and the desire to improve its

performance, given the previously explained benefits that this approach can bring, if compared to ANNs.

Conclusions

This thesis has provided a clear overview of AI applications in construction project cost estimation, shedding light on the functioning, strengths, and limitations of various AI techniques, especially Machine Learning with Artificial Neural Networks, Knowledge-Based Systems with Case-Based Reasoning, and Evolutionary Systems with Genetic Algorithms.

The research indicates that Artificial Neural Networks are at the forefront of AI-based cost estimation, either in their pure form or as the base of a hybrid algorithm (Hashemi et al., 2020). They are widely studied due to their adaptability and their ability to identify relationships with the variables, major forces behind their superior estimating performances. Conversely, they present significant sensitivity to input data and weights, along with the number of hidden layers and neurons. The increasing number of hybrid models in research can be then explained to mitigate these issues. Nevertheless, despite their predictive power, ANNs' black box nature, computational demands, and training complexity are significant drawbacks.

Knowledge-based systems, specifically CBR, have also been revealed very promising in cost estimation (Hu et al., 2016), providing accurate estimates at the early stages of projects and being less time-consuming than traditional methods. Despite their inferior accuracy, their knowledge base allows implicit knowledge to be converted into explicit data, providing a lasting knowledge repository that can be employed for future estimates. Moreover, transparency and the ability to revise results make KBS an attractive alternative.

The last group of systems analyzed are Evolutionary Systems, primarily Genetic algorithms. They are frequently used for optimization purposes (Hashemi et al., 2020) to enhance accuracy and mitigate the limitations of individual algorithms. GA has shown promising results in attribute weight optimization for CBR models as well as ANN models.

Future research in this field should consider the impact of these AI techniques in active construction projects. Indeed, despite the focus of research, there is a lack of reports of field-

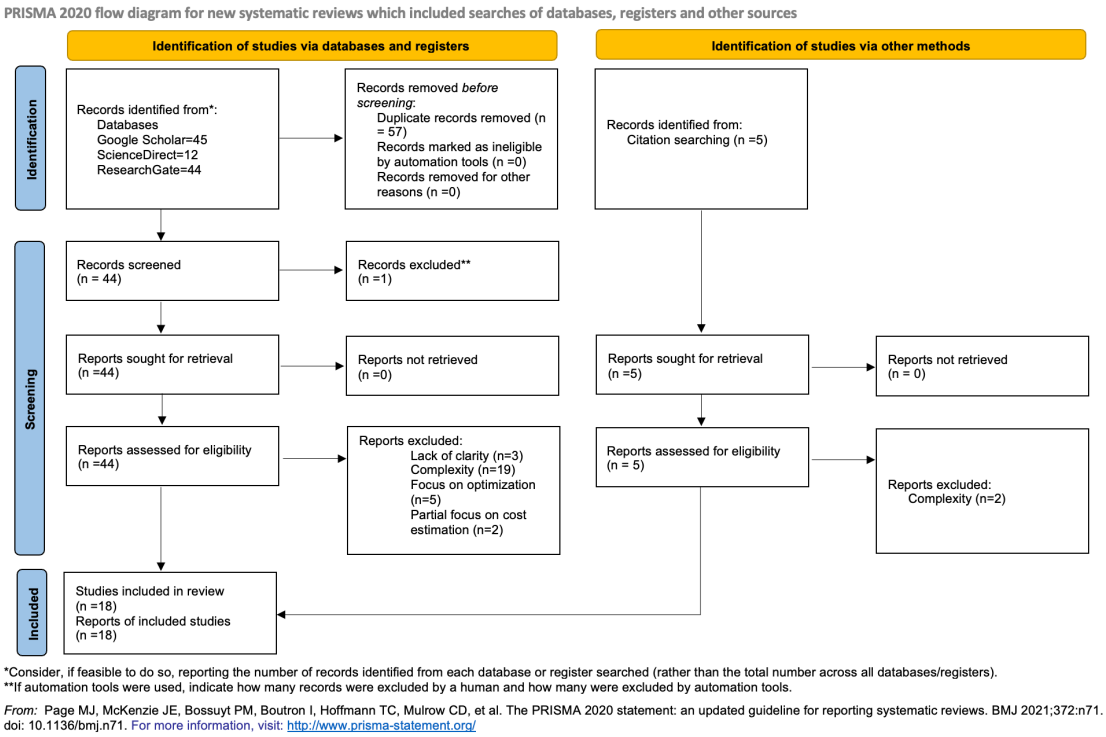
applications of these techniques. Furthermore, exploring the transparency and interpretability of AI models remains a significant challenge, along with their sensitivity to their indexes and structure, which at the moment can be seen as their major obstacle to widespread implementation, given that there is still not a widely appliable structure that guarantees an accurate performance, with researchers still fine-tuning the models to the specific data.

In conclusion, the application of AI techniques in construction project cost estimation aligns with the industry's demand for better project management (Nieto-Rodriguez & Viana Vargas, 2023). The cost of inaccurate estimates, in terms of project delays and cost overruns, highlights the urgency of new methods and AI-based ones could provide the solution. As AI continues to evolve and gain popularity, it is expected to transform project management (Nieto-Rodriguez & Viana Vargas, 2023), enabling project managers to focus on higher-value tasks and improving project success rates. Further research and practical implementation of AI techniques in construction projects are hence essential to harness the full potential of these technologies. Indeed, while algorithms such as ANNs and CBR sport the potential to revolutionize cost estimation in construction projects by offering improved accuracy and adaptability, they still display a high sensitivity to their structure and training, with results widely changing according to the indexing and feature selection.

Appendix

Figure A.1

PRISMA Flow Diagram



Note The diagram shows the selection process for the studies evaluated and employed for this thesis (Page et al., 2021).

Figure A.2

Engineering categories with impact factors in Cheng et al.'s study

Engineering	Features	Impact factors	Values or units
Temporary construction	QT	1. Site area	Meter2
	QT	2. Floors underground	Floors
	QT	3. Floors aboveground	Floors
	QT	4. Total floor area	Meter2
Geotechnical construction	QT	1. Site area	Meter2
	QT	2. Excavation depth	Meter
	QT	3. Floors underground	Floors
	QT	4. Households in adjacent buildings	Households
	QL	5. Soil condition	Stiff, medium, soft
	QL	6. Bracing system	Tied-back, Inside bracing
	QL	7. Retaining structure	None, sheet-pile, soldier pile, rail pile, diaphragm wall, others
Structural construction	QT	1. Total floor area	Meter2
	QT	2. Floors underground	Floors
	QT	3. Floors aboveground	Floors
	QT	4. Area of exterior wall	Meter2
	QL	5. Seismic zone	Type A, B
	QL	6. Soil condition	Stiff, medium, soft
	QL	7. Type of foundations	Raft, pile
	QL	8. Type of Excavation	Partial-braced, top-down, bottom-up, slope excavation
Decorative construction	QT	1. Total floor area	Meter2
	QT	2. Area of exterior wall	Meter2
	QT	3. Households planned	Households
	QT	4. Type of flooring	Ceramic tile, archaized brick, quartz tile, terrazzo tile, wooden, granite tile
	QT	5. Type of ceiling	Emulsion paint, light rigid frame, waterproof, wood board, calcium silicate board, metal
	QT	6. Interior wall decoration	Emulsion paint, ceramic tile, granite tile
	QT	7. Exterior wall decoration	Strip tile, facial cut terrazzo, facial washed terrazzo, granite tile, curtain wall, cast plate
	QT	8. Material of doors	Wooden, aluminum, copper vitriol, stainless steel, fireproof
	QT	9. Material of Windows	Aluminum, plastic-steel, airtight, stainless steel
Electro-mechanical infrastructure	QT	1. Total floor area	Meter2
	QT	2. Households planned	Households
	QT	3. Elevators	Number
	QL	4. Air conditioner	Non-central, central
	QL	5. Kitchen	Luxurious, common, basic
	QL	6. Shower room	Luxurious, common, basic
	QL	7. Fire control	Common, basic
	QL	8. Parking	Mechanic parking system, parking lot
Miscellaneous construction	QT	1. Site area	Meter2
	QT	2. Total floor area	Meter2
	QT	3. Households planned	Households
	QT	4. Floors underground	Floors
	QT	5. Floors aboveground	Floors
Indirect construction	QT	1. Total floor area	Meter2
	QT	2. Floors underground	Floors
	QT	3. Floors aboveground	Floors
	QL	4. Type of excavation	Partial-braced, top-down, bottom-up, slope excavation

Notations: QT – quantitative factor; QL – qualitative factor.

Note The table displays the list of engineering categories of the EFHNN model for the Cheng et al. study (Cheng et al., 2010).

Figure A.3

Correlation factors between input variables and the dependent variable in Xu et al.'s study

P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18
0.378	0.226	0.456	0.242	0.099	0.464	0.068	0.310	0.216	0.371	0.071	0.082	0.221	0.188	0.186	0.187	0.368	0.309

Note The table lists, the correlation indexes between the input variables, listed as follows:

P1: Construction area,

P2: Foundation type,

P3: Foundation bottom elevation,

P4: Structure type,

P5: Seismic intensity,

P6: Roof waterproof,

P7: Roof insulation,

P8: Façade,

P9: Wall insulation,

P10: Windows and doors,

P11: Ceiling, number of stories,

P12: Ground floor,

P13: Wall insulation,

P14: Number of floors,

P15: Floor height,

P16: Top of wall level,

P17: Concrete price,

P18: Steel price (Xu et al., 2015).

Table A.4

List of studies investigated in the review

Name	**Author**	**Year**	**Algorithm**	**Notes**
'Cost estimation during design step: Parametric method versus case-based reasoning method'	Duverlie, Castelain	1999	CBR	Included
'Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning'	Kim, An, Kang	2004	ANN, CBR, MRA	Included

'*Neural network modeling of highway construction costs'*	Wilmot, Mei	2005	ANN	Included
'Determining Attribute Weights in a CBR Model for Early Cost Prediction of Structural Systems'	Doğan, Arditi, Günaydin	2006	CBR	Included
'A case-based reasoning cost estimating model using experience by analytic hierarchy process'	An, Kim, Kang	2007	CBR and AHP	Included
'Web-based conceptual cost estimates for construction projects using Evolutionary Fuzzy Neural Inference Model'	Cheng, Tsai, Sudjono	2009	Fuzzy NN	Included
'Conceptual cost estimates using evolutionary fuzzy hybrid neural network for projects in the construction industry'	Cheng, Tsai, Sudjono	2010	Fuzzy NN, GA	Included
'A CBR-based hybrid model for predicting a construction duration and cost based on project characteristics in multi-family housing projects'	Koo, Hong, Hyun, Koo	2010	CBR	Included
'Preliminary Cost Estimation Model Using Case-Based Reasoning and Genetic Algorithms'	Kim, Kim	2010	CBR, GA	Included
'CBR Revision Model for Improving Cost Prediction	Ji, Hyun, Hong	2010	CBR	Included

Accuracy in Multifamily Housing Projects'				
'Early stage cost estimation of buildings construction projects using artificial neural networks'	Arafa, Alqedra	2011	ANN	Not Included Complexity
'A Neural Network Model for Building Construction Projects Cost Estimating'	El-Sawalhi, Shehatto	2014	ANN	Not included Complexity
'Using Intelligent Techniques in Construction Project Cost Estimation: 10-Year Survey'	Elfaki, Alatawi, Abushandi	2014	Review	Not Included Redundant
'Conceptual Cost Estimation Model for Engineering Services in Public Construction Projects'	Hyari, Al-Daraiseh, El-Mashaleh	2015	ANN	Not Included Complexity
'The Case-based Reasoning Model of Cost Estimation at the Preliminary Stage of a Construction Project'	Zima	2015	CBR	Included
'Construction Project Cost Prediction Based on Genetic Algorithm and Least Squares Support Vector Machine'	Xu, Xu, Zhou, Wu	2015	SVM, GA	Included
'The application of case-based reasoning in construction management research: An overview'	Hu, Xia, Chen	2016	Review	Included
'Estimation of Costs and Durations of Construction of	Peško, MuIenski, Šešlija, Radović,	2017	ANN, SVM	Not Included Complexity

Urban Roads Using ANN and SVM'	Vujkov, Bibić, Krklješ			
'Cost Calculation of Construction Projects Including Sustainability Factors Using the Case-Based Reasoning (CBR) Method'	Lesniak, Zima	2018	CBR	Not Included Complexity
'ANN Based Approach for Estimation of Construction Costs of Sports Fields'	Juszczyk, Lesniak, Zima	2018	ANN	Included
'Enhanced Predictive Models for Construction Costs: A Case Study of Turkish Mass Housing Sector'	Ugur, Kanit, Erdal, Namli, Erdal, Baykan Erdal	2019	ANN (MLP), CART	Not Included Complexity
'Investigating profitability performance of construction projects using big data: A project analytics approach'	Bilal, Oyedele, Kusimo, Owolabi, Akanbi, Ajayi, Akinade, Delgado	2019	Hybrid KBS model	Not Included Optimization
'Cost estimation and prediction in construction projects: a systematic review on machine learning techniques'	Hashemi, Ebadati, Kaur	2020	Review	Included
'*Performance evaluation of normalization based CBR models for improving construction cost estimation'*	Ahn, Ahn, Ji	2020	CBR	Not Included Complexity

'Construction Cost Estimation Using a Case-Based Reasoning Hybrid Genetic Algorithm Based on Local Search Method'	Jung, Pyeon, Park, Lee, Yoon, Rho	2020	CBR, GA	Not Included Complexity
'Integrated Approach of Cost Estimation for Road Projects by Using Artificial Neural Network (ANN) and Analytic Hierarchy Process (AHP)'	Teferi, Debela, Tadesse	2021	ANN, AHP	Not Included Complexity
'Develop an artificial neural network (ANN) model to predict construction projects performance in Syria'	Maya, Hassan, Hassan	2021	ANN	Not Included Clarity
'Developing an Integrative Data Intelligence Model for Construction Cost Estimation'	Ali, Burhan, Kassim, Al-Khafaji	2022	ANN, XG Boost, RF, SVM	Not Included Complexity
'Conceptual estimation of construction duration and cost of public highway projects'	Mohamed, Moselhi	2022	ANN, SVM, RF	Included
'Extreme Gradient Boosting-Based Machine Learning Approach for Green Building Cost Prediction'	Alshboul, Shehadeh, Almasabha, Almuflih	2022	XG Boost, DNN, RF	Not Included Complexity
'Developing six hybrid machine learning models based on Gaussian process regression and meta-heuristic optimization algorithms for prediction of	Mahmoodzadeh, Nejati, Mohammadi, Ibrahim, Khishe, Rashidi, Mohammed	2022	GPR	Not Included Complexity

duration and cost of road tunnels construction'				
'Machine learning-based cost predictive model for better operating expenditure estimations of U.S. light rail transit projects'	Zhou, Etemadi, Mardon	2022	ANN, MLR	Not Included Partial focus on costs
'Optimized machine learning modeling for predicting the construction cost and duration of tunnelling projects'	Mahmoodzadeh, Nejati, Mohammadi,	2022	GPT, DT, SVM, LR	Not Included Optimization
'A Machine Learning Study to Enhance Project Cost Forecasting'	Inan, Narbaev, Hazir	2022	ANN, EVM	Not Included Complexity
'Artificial Intelligence in Construction Projects: A Systematic Scoping Review'	Bang, Olsson	2022	Review	Not Included Partial focus on costs
'Forecasting Construction Price Index using Artificial Intelligence Models: Support Vector Machines and Radial Basis Function Neural Network'	Nguyen T.T., Nguyen D.D., Nguyen S.D. Nguyen, Prakash, Tran	2022	SVM, RBFN	Not Included Complexity
'Predictive Analytics for Early-Stage Construction Costs Estimation'	Miranda, Castillo, Gonzalez, Adafin	2022	Review	Included

'Bootstrap Aggregated Case-Based Reasoning Method for Conceptual Cost Estimation'	Uysal, Sonmez	2023	CBR	Not Included Complexity
'An artificial neural network (ANN) approach for early cost estimation of concrete bridge systems in developing countries: the case of Sri Lanka'	Fernando, Zhang, Dilshan	2023	ANN	Not Included Clarity
'The Adoption of a Machine Learning Approach in a Big Data Concept to Predict Project Cost Budgeting in the Thai Auction Process of Procurement Management for a Construction Project'	Kusonkhum, Srinavin, Chaitongrat	2023	ANN, DT, K-NN	Not Included Complexity
'Early Highway Construction Cost Estimation: Selection of Key Cost Drivers'	Simić, Ivanišević, Nedeljković, Senić, Stojadinović, Ivanović	2023	XG Boost, ANN, MRA	Not Included Optimization
'Enhancing Accuracy in Cost Estimation for Façade Works: Integration of Case Based Reasoning, Random Forest, and Artificial Neural Network Techniques'	Long, Anh	2023	CBR, ANN, RF	Not Included Clarity
'Intelligent Analysis of Construction Costs of Shield Tunneling in Complex	Ye, Zhou, Ding, Jin	2023	RF, K-NN	Not Included Optimization

Geological Conditions by Machine Learning Method'				
'An artificial neural network (ANN) approach for early cost estimation of concrete bridge systems in developing countries: the case of Sri Lanka'	Fernando, Dilshan, Zhang	2023	ANN	Not Included Complexity
'Assessing effects of economic factors on construction cost estimation using deep neural networks'	Wang, Cheung, Asghari, Hsu	2021	DNN	Not Included Complexity
'A new model for cost estimation construction project using Hybrid importance regression ensemble method'	Alkhuadhan, Naimi	2023	IERA, K-NN	Not Included Complexity
'A machine learning study to improve the reliability of project cost estimates'	Narbaev, Hazir, Kamitova, Talgat	2023	XG Boost, EVM	Not Included Complexity
'Artificial Neural Networks for Cost Estimation of Road Projects in Nepal'	Acharya, Karki	2023	ANN	Not Included Complexity
'Performance evaluation of normalization based CBR models for improving construction cost estimation'	Ahn, Ji, Kim	2023	CBR	Not included Optimization
'Causes and Effects of Cost Overruns in Construction Projects'	Atapattu, Sutrisna, Domingo	2023	Review	Included

Figure A.5

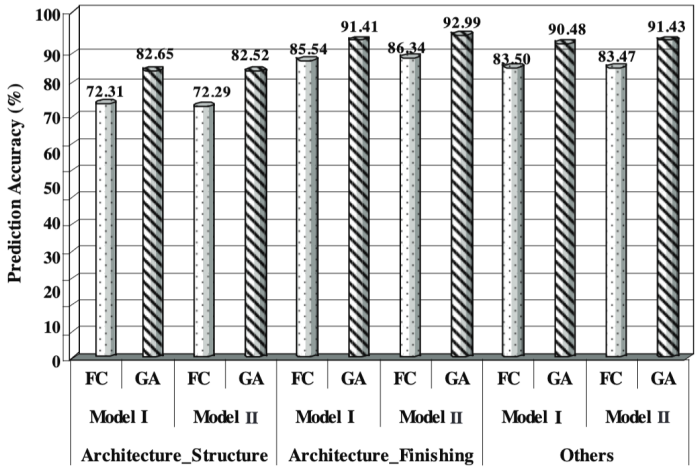
Optimized weights and average error of the algorithm's output in Doğan et al.'s study.

Weight generation method	Attribute weights								Average error in CBR prediction (%)
	Total area	Ratio of floor area to total area	Ratio of footprint area to total area	Number of floors	Overhang design	Core location	Floor type	Foundation system	
Feature counting	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	17.63
Gradient descent	0.0069	0.1885	0.1528	0.1427	0.1049	0.1560	0.0316	0.2161	21.20
Genetic algorithms	1.0000	2.0056	1.0010	9.9988	1.0031	1.0000	3.9999	1.0000	16.23

Note The table depicts the accuracy of the three optimization methods employed in the retrieval phase of Doğan et al.'s cost estimation study (Doğan et al., 2006).

Figure A.6

Accuracy differences with different optimization algorithms



Note The histogram reports the significant differences between an FC and GA-based optimization in the structural and finishing estimations for a construction project (Koo et al., 2010).

Resources

Abourizk, S. M., Babey, G. M., & Karumanasseri, G. (2002). Estimating the cost of capital projects: an empirical study of accuracy levels for municipal government projects. *Canadian Journal of Civil Engineering*, *29*(5), 653–661. https://doi.org/10.1139/l02-046

Alexander, L. K., Lopes, B., Ricchetti-Masterson, K., & Yeatts, K. B. (n.d.). *Sources of Systematic Error or Bias: Information Bias*.

Altenberg, L. (2016). Evolutionary Computation. In *Encyclopedia of Evolutionary Biology* (pp. 40–47). Elsevier. https://doi.org/10.1016/B978-0-12-800049-6.00307-3

An, S.-H., Kim, G.-H., & Kang, K.-I. (2007). A case-based reasoning cost estimating model using experience by analytic hierarchy process. *Building and Environment*, *42*(7), 2573–2579. https://doi.org/10.1016/j.buildenv.2006.06.007

Approaches to Cost Estimation. (2018). Cotney Attorneys & Consultants. https://www.cotneycl.com/approaches-to-cost-estimation/

Artificial Intelligence market size/revenue comparisons 2022 . (n.d.). Statista. Retrieved October 29, 2023, from https://www.statista.com/statistics/941835/artificial-intelligence-market-size-revenue-comparisons/

Atapattu, C., Domingo, N., & Sutrisna, M. (2023). *Causes and Effects of Cost Overruns in Construction Projects*. https://www.researchgate.net/publication/366897935

Availability Heuristic. (n.d.). APA Dictionary of Psychology. Retrieved April 28, 2023, from https://dictionary.apa.org/availability-heuristic

Bai, Z., Wei, G., Liu, X., & Zhao, W. (2014). *Predictive Model of Energy Cost in Steelmaking Process Based on BP Neural Network*. https://doi.org/10.2991/sekeie-14.2014.18

Basak, S. K. (2018, July 5). *How to perform Roulette wheel and Rank based selection in a genetic algorithm?* |. Medium. https://setu677.medium.com/how-to-perform-roulette-wheel-and-rank-based-selection-in-a-genetic-algorithm-d0829a37a189

Behera, S., Sarat, ·, Nayak, C., & Pavan Kumar, · A V S. (2023). A Comprehensive Survey on Higher Order Neural Networks and Evolutionary Optimization Learning Algorithms in Financial Time Series Forecasting. *Archives of Computational Methods in Engineering*, *30*, 4401–4448. https://doi.org/10.1007/s11831-023-09942-9

Belack Carl. (2022). *Budget and overhead*.

Borowicz, J. J., Psp, C., Brown, R. B., Donaldson, D. C., Dysert, L. R., Cep, C., Garcia Da Roza, R., Hollman, J. M., Hollmann, J. K., Cep, C., Kutilek, F., Mccuen, T. L., Parker, D. E., Todd, C., Pickett, W., Sinnathamby, K., Stephenson, C. H. L., Uppal, K. B., & Whiteside, J. D. (2020). *Cost estimate classification system-As applied in engineering, procurement, and construction for the building and general construction industries*.

Burek, P. (2008). Creating clear project requirements: differentiating ``what'' from ``how'' Paper presented at PMI\textregistered{} Global Congress. *North America*.

Butler, K. (n.d.). *InfoGuides: Systematic Reviews: Prepare Protocol, PICO and PRISMA flow chart*. Retrieved October 29, 2023, from https://infoguides.gmu.edu/SR/prepare

Casad, B. J., & Luebering, J. E. (2023, February 3). *Confirmation bias | Definition, Examples, Psychology, & Facts | Britannica*. Encyclopedia Britannica. https://www.britannica.com/science/confirmation-bias

Castro Miranda, S. L., Del Rey Castillo, E., Gonzalez, V., & Adafin, J. (2022). Predictive Analytics for Early-Stage Construction Costs Estimation. *Buildings*, *12*(7), 1043. https://doi.org/10.3390/buildings12071043

Cheng, M.-Y., Tsai, H.-C., & Hsieh, W.-S. (2009). Web-based conceptual cost estimates for construction projects using Evolutionary Fuzzy Neural Inference Model. *Automation in Construction*, *18*(2), 164–172. https://doi.org/10.1016/j.autcon.2008.07.001

Cheng, M.-Y., Tsai, H.-C., & Sudjono, E. (2010). Conceptual cost estimates using evolutionary fuzzy hybrid neural network for projects in construction industry. *Expert Systems with Applications*, *37*(6), 4224–4231. https://doi.org/10.1016/j.eswa.2009.11.080

Christensen, P., Larry Dysert, C. R., Jennifer Bates, C., Dorothy Burton Robert C Creese, C. J., CCE John Hollmann, P. K., CCE Kenneth Humphreys, P. K., CCE Donald McDonald, P. F., CCE C Arthur Miller Bernard A Pietlock, J. P., & Wesley Querns, C. R. (1997). *17R-97: Cost Estimate Classification System*.

Classification confusion matrix. (n.d.). MATLAB. Retrieved May 2, 2023, from https://www.mathworks.com/help/deeplearning/ref/confusion.html

Corne, D., & Lones, M. A. (2018). Evolutionary Algorithms. In *Handbook of Heuristics* (pp. 409–430). Springer International Publishing. https://doi.org/10.1007/978-3-319-07124-4_27

Corzo Cynthia. (2022). *Overusing Email for Certain Conversations Affects Later Performance - FIU Business Now Magazine*. FIU BUSINESS NOW. https://business.fiu.edu/news/2022/overusing-email-for-certain-conversations-affects-later-performance.html

Costello, K. (Ed.). (2019). Project Management Tasks to Be Eliminated as AI Takes Over. In *Gartner Program & Portfolio Management Summit*. https://www.gartner.com/en/newsroom/press-releases/2019-03-20-gartner-says-80-percent-of-today-s-project-management

Cote, C. (2021, December 14). *What Is Regression Analysis in Business Analytics?* Harvard Business School. https://online.hbs.edu/blog/post/what-is-regression-analysis

Definition of Case-based Reasoning (CBR). (n.d.). Gartner Information Technology Glossary. Retrieved September 21, 2023, from https://www.gartner.com/en/information-technology/glossary/cbr-case-based-reasoning

Doğan, S. Z., Arditi, D., Asce, M., & Murat Günaydın, H. (2006). *Determining Attribute Weights in a CBR Model for Early Cost Prediction of Structural Systems*. https://doi.org/10.1061/ASCE0733-93642006132:101092

Duverlie, P., & Castelain, J. M. (1999). Cost estimation during design step: Parametric method versus case based reasoning method. *International Journal of Advanced Manufacturing Technology*, *15*(12), 895–906. https://doi.org/10.1007/S001700050147/METRICS

Eiben, A. E., & Smith, J. (2015). From evolutionary computation to the evolution of things. *Nature*, *521*(7553), 476–482. https://doi.org/10.1038/nature14544

Equipment Cost Setup | Druml Group, Inc. (n.d.). Retrieved April 24, 2023, from https://www.druml.com/management-advisory/job-cost-control/equipment-cost-setup/

Fuzzy Rules. (n.d.). ScienceDirect. Retrieved May 13, 2023, from https://www.sciencedirect.com/topics/chemical-engineering/fuzzy-rules

Gandhi, R. (2018, July 7). *Support Vector Machine — Introduction to Machine Learning Algorithms*. Towards Data Science. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. http://oreilly.com

Gills, A., & Moore, J. (2023, March 28). *What is a Knowledge-based System? | Definition from TechTarget*. https://www.techtarget.com/searchcio/definition/knowledge-based-systems-KBS

Global Construction Market Report And Strategies To 2032. (2023, October). The Business Research Company. https://www.thebusinessresearchcompany.com/report/construction-market

Greiman, Virginia. (2013). *Megaprojects: lessons on risk and project management from the Big Dig*.

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

Hastie, S., & Wojewoda, S. (2015). *Standish Group 2015 Chaos Report*. https://www.infoq.com/articles/standish-chaos-2015/

Hegazy, T. (2002). *Computer-based construction project management*. Pearson.

Hendrickson, C. (2008). Cost Estimation. In *Project Management for Construction* (2nd ed.). Carnegie Mellon University. https://www.cmu.edu/cee/projects/PMbook/05_Cost_Estimation.html

Herszon, L., & Keraminiyage, K. (2014). Dimensions of project complexity and their impact on cost estimation. Paper presented at PMI\textregistered{} Global Congress. *North America*.

Hsu, K., Gupta, H. V., & Sorooshian, S. (1995). Artificial Neural Network Modeling of the Rainfall-Runoff Process. *Water Resources Research*, *31*(10), 2517–2530. https://doi.org/10.1029/95WR01955

Hu, X., Xia, B., Skitmore, M., & Chen, Q. (2016). The application of case-based reasoning in construction management research: An overview. *Automation in Construction*, *72*, 65–74. https://doi.org/10.1016/j.autcon.2016.08.023

Hyari, K. H., Al-Daraiseh, A., & El-Mashaleh, M. (2016). Conceptual Cost Estimation Model for Engineering Services in Public Construction Projects. *Journal of Management in Engineering*, *32*(1). https://doi.org/10.1061/(ASCE)ME.1943-5479.0000381

Hyperbolic tangent sigmoid transfer function. (n.d.). MATLAB. Retrieved May 2, 2023, from https://www.mathworks.com/help/deeplearning/ref/tansig.html

Ille R ', B. L. M., & Goldberg, D. E. (1995). Genetic Algorithms , Tournament Selection, and the Effects of Noise. In *Com plex Sy* (Vol. 9).

Ji, C., Hong, T., & Hyun, C. (2010). *A CBR Revision Model for Improving Cost Prediction Accuracy in Multifamily Housing Projects*. https://doi.org/10.1061/ASCEME.1943-5479.0000018

Jordan. (2021). RFP experts learning center: Does the lowest bid win the contract? In *The Bid Lab*.

Juszczyk, M., Leśniak, A., & Zima, K. (2018). ANN Based Approach for Estimation of Construction Costs of Sports Fields. *Complexity*, *2018*, 1–11. https://doi.org/10.1155/2018/7952434

Kaya, Y., Uyar, M., & Tekđn, R. (2011). *A Novel Crossover Operator for Genetic Algorithms: Ring Crossover*. https://www.researchgate.net/publication/220485962

Kim, G.H., An, S.H., & Kang, K.I. (2004). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. *Building and Environment*, *39*(10), 1235–1242. https://doi.org/10.1016/j.buildenv.2004.02.013

Kim, K. J., & Kim, K. (2010). Preliminary Cost Estimation Model Using Case-Based Reasoning and Genetic Algorithms. *Journal of Computing in Civil Engineering*, *24*(6), 499–505. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000054

Koo, C. W., Hong, T. H., Hyun, C. T., Park, S. H., & Seo, J. O. (2010). A study on the development of a cost model based on the owner's decision making at the early stages of a construction project. *International Journal of Strategic Property Management*, *14*(2), 121–137. https://doi.org/10.3846/ijspm.2010.10

Leach, A. R. (2006). Ligand-based approaches: Core molecular modeling. *Comprehensive Medicinal Chemistry II*, *4*, 87–118. https://doi.org/10.1016/B0-08-045044-X/00246-7

Lovallo, D., & Kahneman, D. (2003a). Delusions of success: How optimism undermines executives' decisions. *Harv. Bus. Rev.*

Ludden, M. (2019). *How AI will Transform Project Management*. https://graduate.northeastern.edu/resources/ai-and-project-management/

Machine Learning Algorithms. (n.d.). Microsoft Azure. Retrieved April 29, 2023, from https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-are-machine-learning-algorithms/

Mahtani, K., Spencer, E. A., Brassey, J., & Heneghan, C. (2018). Catalogue of bias: observer bias. *BMJ Evidence-Based Medicine*, *23*(1), 23–24. https://doi.org/10.1136/ebmed-2017-110884

Marr, B. (2016, February 19). *A Short History of Machine Learning*. Forbes. https://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learning-every-manager-should-read/?sh=75e2d0fb15e7

Matt Plummer. (2019, January 22). *How to Spend Way Less Time on Email Every Day*. Harvard Business Review. https://hbr.org/2019/01/how-to-spend-way-less-time-on-email-every-day

Menaga, D., & Saravanan, S. (2021). Application of artificial intelligence in the perspective of data mining. In *Artificial Intelligence in Data Mining* (pp. 133–154). Elsevier. https://doi.org/10.1016/B978-0-12-820601-0.00006-9

Messner, J. (2019). *Chapter 5: Introduction to Construction Cost Estimating*. The Pennsylvania State University.

Mohamed, B., & Moselhi, O. (2022). Conceptual estimation of construction duration and cost of public highway projects. *J. Inf. Technol. Constr.*, *27*, 595–618.

Nauck, D., & Kruse, R. (1999). Neuro-fuzzy systems for function approximation. *Fuzzy Sets and Systems*, *101*(2), 261–271. https://doi.org/10.1016/S0165-0114(98)00169-9

Nieto-Rodriguez, A., & Viana Vargas, R. (2023, February 2). *How AI Will Transform Project Management*. Harvard Business Review. https://hbr.org/2023/02/how-ai-will-transform-project-management

NIGPT-The institute for public procurement. (n.d.). *Dictrionary of Terms*. Retrieved April 24, 2023, from https://www.nigp.org/dictionary-of-terms?search=responsive&page=1

OpenAI. (2023, October 30). *Concept Notes*. https://chat.openai.com/share/1c977d73-3d5b-4cc4-a643-1517f30ce670

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, n71. https://doi.org/10.1136/bmj.n71

PMI. (2013). *Guide to the Project Management Body of Knowledge*. Project Management Institute.

PMOtto.ai – Ricardo Viana Vargas. (n.d.). Retrieved April 27, 2023, from https://ricardo-vargas.com/special-projects/pmotto/

PMP, & Ralph L. Kliem. (2007). *Effective Communications for Project Management*.

Pray, J., & Walsh, K. (n.d.). Managing the impacts of material price fluctuations on construction projects. In *JD Supra*.

Ramadhan, L. (2021, November 10). *Radial Basis Function Neural Network Simplified*. Towards Data Science. https://towardsdatascience.com/radial-basis-function-neural-network-simplified-6f26e3d5e04d

Ribeirinho, M. J., Mischke, J., Sjödin, E., Biörck, J., Anderson, T., Blanco, J. L., Palter, R., & Rockhill, D. (2020, June 4). *The next normal in construction*. https://www.mckinsey.com/capabilities/operations/our-insights/the-next-normal-in-construction-how-disruption-is-reshaping-the-worlds-largest-ecosystem#/

Samui, P., & Kothari, D. P. (2011). Utilization of a least square support vector machine (LSSVM) for slope stability analysis. *Scientia Iranica*, *18*(1), 53–58. https://doi.org/10.1016/j.scient.2011.03.007

Sanni, S. E., Okoro, E. E., Sadiku, E. R., & Oni, B. A. (2022). Advances in data-centric intelligent systems for air quality monitoring, assessment, and control. *Current Trends and Advances in*

Computer-Aided Intelligent Environmental Data Engineering, 25–58. https://doi.org/10.1016/B978-0-323-85597-6.00021-5

Santosh, K., Das, N., & Ghosh, S. (2022). Deep Learning Models for Medical Imaging. In *Deep Learning Models for Medical Imaging* (Vol. 1, pp. 1–27). Elsevier. https://doi.org/10.1016/B978-0-12-823504-1.00011-8

Scholl-Sternberg, A. (n.d.). *Crossrail Integration Facility and Test Automation-improving resilience with automated testing*.

Sharma, S., Ahmed, S., Naseem, M., Alnumay, W. S., Singh, S., & Cho, G. H. (2021). A Survey on Applications of Artificial Intelligence for Pre-Parametric Project Cost and Soil Shear-Strength Estimation in Construction and Geotechnical Engineering. *Sensors*, *21*(2), 463. https://doi.org/10.3390/s21020463

Soni, D. (2018, February 18). *Introduction to Evolutionary Algorithms*. Towards Data Science. https://towardsdatascience.com/introduction-to-evolutionary-algorithms-a8594b484ac

The Big Dig: project background. (n.d.). Mass.Gov. Retrieved November 10, 2023, from https://www.mass.gov/info-details/the-big-dig-project-background

Thevenot, A. (2020, December 21). *Particle Swarm Optimization (PSO) Visually Explained* . Towards Data Science. https://towardsdatascience.com/particle-swarm-optimization-visually-explained-46289eeb2e14

Tkáč, M., & Verner, R. (2016). Artificial neural networks in business: Two decades of research. *Applied Soft Computing*, *38*, 788–804. https://doi.org/10.1016/j.asoc.2015.09.040

Turner, J. R., & Cochrane, R. A. (1993). Goals-and-methods matrix: coping with projects with ill defined goals and/or methods of achieving them. *International Journal of Project Management*, *11*(2), 93–102. https://doi.org/10.1016/0263-7863(93)90017-H

U.S. Government Accountability Office. (n.d.). *Cost estimating and assessment guide: Best practices for developing and managing program costs*.

Using ENR Indexes. (n.d.). Engineering News-Record. Retrieved September 27, 2023, from https://www.enr.com/economics/faq

Usmani, F. (2022). Contingency reserve vs management reserve. In *PMP*. Fahad Usmani.

What are Neural Networks? (n.d.). IBM. Retrieved November 4, 2023, from https://www.ibm.com/topics/neural-networks

What is a Decision Tree. (n.d.). IBM. Retrieved May 12, 2023, from https://www.ibm.com/topics/decision-trees

What is complexity in project management? (n.d.). ICCPM. Retrieved April 24, 2023, from https://iccpm.com/about-complex-project-management/

What is evolutionary computation? (n.d.). TechTarget. Retrieved September 25, 2023, from https://www.techtarget.com/whatis/definition/evolutionary-computation

What is Gradient Descent? . (n.d.). IBM. Retrieved November 10, 2023, from https://www.ibm.com/topics/gradient-descent

What is Machine Learning? (n.d.-a). IBM. Retrieved April 29, 2023, from https://www.ibm.com/topics/machine-learning

What is Machine Learning? (n.d.-b). MATLAB. Retrieved April 29, 2023, from https://www.mathworks.com/discovery/machine-learning.html

What is Overfitting? (n.d.). IBM. Retrieved May 12, 2023, from https://www.ibm.com/topics/overfitting

What is Random Forest? (n.d.). IBM. Retrieved May 12, 2023, from https://www.ibm.com/topics/random-forest

What Is Roulette Wheel Selection in Genetic Algorithms? (n.d.). Saturn Cloud Blog. Retrieved September 27, 2023, from https://saturncloud.io/blog/what-is-roulette-wheel-selection-in-genetic-algorithms/

Wilmot, C. G., & Mei, B. (2005). Neural network modeling of highway construction costs. *Journal of Construction Engineering and Management*, *131*, 765–771.

Xu, M., Xu, B., Zhou, L., & Wu, L. (2015). *Construction Project Cost Prediction Based on Genetic Algorithm and Least Squares Support Vector Machine*.

Zima, K. (2015). The Case-based Reasoning Model of Cost Estimation at the Preliminary Stage of a Construction Project. *Procedia Engineering*, *122*, 57–64. https://doi.org/10.1016/j.proeng.2015.10.007