Predicting construction costs in the program phase of the construction process: a machine learning approach

Jaap Beltman

Supervisor: Rogier Harmelink
2nd assessor: Engin Topan
Defence date: 31/08/2021
Industrial Engineering & Management thesis

UNIVERSITY OF TWENTE.

ARCADIS

Preface

Dear reader,

This bachelor thesis finalizes my bachelor's program in Industrial Engineering & Management. This thesis was executed at Arcadis at the Cost & Data Management department.

Last year was not a regular year with the COVID-19 pandemic. It required flexibility from everyone involved in this thesis and I am grateful for the result. I learned a lot at Arcadis and enjoyed the time there. I am glad I challenged myself in this bachelor thesis with subjects I was not familiar with before. It gave me new insights, that were particularly useful for selecting my master's program.

I want to thank Sander van Gemert, Jeroen Hartman, Rogier Harmelink, Engin Topan and Ipek Seyran Topan for the trust and guidance. Also, graduating at Arcadis was a great experience and the colleagues kept me motivated while doing my bachelor's thesis.

Jaap Beltman
Enschede, Juli 2021

Abstract

Cost estimation in the program phase of the construction process can only be roughly estimated at the moment of writing. On the other hand, the use of data will become more important in the future and with sufficient data, statistical predictions can be made using machine learning. Following this trend, machine learning could possibly improve the cost estimation in the program phase of the construction process. The conventional cost estimation method relies on a combination of analogous estimation and expert judgement. Machine learning, which is classified as parametric cost estimation, could initiate a shift in cost estimation methods. Earlier studies that used machine learning to perform cost estimation in the construction sector yielded promising results. The objective for a solution in this thesis is defined as when the created machine learning model could generate predictions within a 20% error range.

Several building quantities have been used as input for the machine learning model. The machine learning types that have been selected are random forest and support vector machine. Two classes of building types have been used to train the model two separate times. Respectively 48% and 46% of all predictions of the two project classes were within a 20% error range. This contradicted the expectations and earlier studies. Evaluation of the model turned out that the model could potentially not predict the cost specific enough.

Contents

List of figures

List of tables

List of equations

1 Introduction

This thesis is written at Arcadis, a global company working with natural and built assets. Section 1.1 elaborates more on the company. The problem of this thesis is explained in section 1.2 and eventually, the research design is set up in section 1.3.

1.1 Company description

Arcadis is the world's leading company delivering sustainable design, engineering and consultancy solutions for natural and built assets (Arcadis, 2021). Arcadis has over 27.000 employees in more than 70 countries, dedicated to improving the quality of life. Arcadis started as the Dutch Heidemaatschappij and is at the moment a rapidly growing company with its headquarters in Amsterdam Zuidas. The five core values for Arcadis are people first, client success, integrity, sustainability and collaboration. These five values construct a basis for all the activities of Arcadis.

In the Cost & Data Management department (C&DM) a broad variety of activities is conducted. C&DM is specialized in strategic advice for policy regarding real estate value, performance, sustainability, money, quality and time. Moreover, the department is specialized in assessing and supervising projects, infrastructure and installations. Also, C&DM investigates and maintains building data for different kinds of buildings.

1.2 Problem identification

This section will sketch the problem context the construction sector, and therefore also Arcadis, faces. The context of the research is stated in section 1.2.1, whereas section 1.2.2 provides the problem statement that is derived from the research context.

1.2.1 Research context

Much of the data in the world has been created in the last few years. The volume of data generated is rapidly growing in the world (Cisco, 2018). According to Agarwal et al. (2016) the construction industry is a sector that generates vast amounts of data, however, the majority of the data is not even processed. Appropriate data processing and management could expose new knowledge and facilitate in responding to emerging opportunities and challenges in a timely manner. Nevertheless, the growth of data in the digital world seems to out-speed the advance of the many existing computing infrastructures (Sivarajah et al., 2016). Therefore, the construction industry could benefit from appropriate data processing.

Information and data are essential components of the success of all organizations operating on the market and in the future will be even more important (Kubina et al., 2015). According to Carande et al. (2017) data & analytics should be the pulse of the organization, incorporated into all key decisions. With appropriate statistical forecasting methods, regular patterns in data can be identified, assuming that sufficient data is available for this purpose (Fildes & Goodwin, 2021). These patterns can be used to support decision-making. To facilitate evidence-based decision-making, organizations need efficient methods to process large volumes of assorted data into meaningful comprehensions (Sivarajah et al., 2016).

The Boston Consultancy Group (2019) points out that it is common for companies to produce forecasts manually, but only a few companies use algorithms. With algorithmically derived forecasting, companies can take action to preempt unfavorable outcomes and promote competitive advantage. In the construction sector, accurate forecasting is possible, but it is a time-consuming process. Algorithms could be less time-intensive than conventional methods.

Forecasting can be performed quantitatively or qualitatively. Qualitative methods draw data from people, organizations, texts, environments, objects and events. Quantitative methods retrieve their data from measuring. The advantage of drawing the data via a quantitative method is that it is suitable for prediction (Cooper & Schindler, 2014). Machine learning is an algorithmic forecasting method that can process quantitative data to make predictions. According to Xu et al. (2021), machine learning has been making major changes in various industries and it has become a powerful tool in the construction sector since it can automize processes. Machine learning technologies play an important role, especially when processing large amounts of data brings a significant added value to saving time and maximizing computing resources. This way, machine learning could be a suitable way for the construction industry to make predictions.

Machine learning is an overarching concept for several methods that include algorithms and improve using data. Cost functions in machine learning can be used to predict costs of all kinds, including construction costs. Xu et al. (2021) in specific points out that in history machine learning is used for costs predictions in the construction sector.

The aim of a construction process is to realize buildings; however, it is often complex and unique. The construction process can be divided into several phases to make the process controllable. The building process follows structured phases. In Dutch context (*Norms and rules according to Normalisatie Instituut*), the building process consist of six stages: *programma*, *ontwerp*, *uitwerking*, *realisering*, *beheer* and *sloop*[1]. In the program stage, fundamental principles are established. Resulting in that in this stage it is already possible to determine the construction costs, however, this is a fast and rough estimation. The deeper in the construction process, how more accurately the construction costs can be predicted since more cost-related information is available. Nevertheless, at the same time calculating construction costs becomes very accurate, resulting in that predicting loses its potential relative advantage of predicting cost accurately. Even though, predicting costs in a later stage of the construction process could save time, (Pijpers & van der Woude, 2012) points out that due to automatization the calculation time becomes small. Therefore, predicting in an early stage of the construction process could be more valuable than predicting in later stages. In the program phase, important information is available, such as the statement of requirements. With this data only very rough cost calculations can be made (Pijpers & van der Woude, 2012). At this point, predicting construction costs accurately could yield an advantage. Regarding the lack of accurate calculations, the program phase stage is a viable stage to apply machine learning.

[1] After this the following translations will be used. *Programma* = program, *ontwerp* = design, *uitwerking* = pre-building phase, *realisering* = realization, *beheer* = maintainance, *sloop* = demolition.

1.2.2 Problem statement

As the generation and collection of data in the world grows, the construction industry participates in this trend. This data can expose new knowledge. However, the growing data collection has outpaced our ability to process the data. The use of data will become more important in the future and with sufficient data, statistical forecasts can be made, exposing patterns in data. Even though it is common for companies to forecast manually, an algorithmic approach can yield a competitive advantage. Machine learning is a forecasting method that can use quantitative data to make algorithmic-based predictions. Moreover, machine learning is a powerful tool for the construction sector and can bring significant added value for saving time and maximizing computing resources. Going deeper into the construction process results in a more accurate prediction of construction costs since more cost-related information is available. However, calculations can already reach a high accuracy in a short time interval, resulting in that machine learning loses its relative potential advantage of predicting construction costs more accurately in a shorter time interval, as can be seen in Table 1. In the program phase cost estimations are not accurate at all and machine learning predictions in this stage could yield accurate costs estimation. An accurate prediction in this early phase using machine learning can result in a new method of estimating cost.

Phase process	Method	Percentage error	Time
Program phase	Key cost figures	Rough estimation	Very quick
Design phase	Element estimation	± 5%	Quick
Design phase	Partial estimation	± 2%	Quick
Pre-building phase	Detailed estimation	± <2%	Quick[2]

Table 1: Cost estimation methods compared, source: (Pijpers & van der Woude, 2012).

[2] Automatization needs to be applied.

1.3 Research design

1.3.1 Research objective

Based on the lack of efficient data processing and the importance of analyzing data, this thesis explored the possibilities of predicting construction costs as accurately as possible in the program phase of the construction process using machine learning. The primary objective of this study was to explore how construction costs predictions with machine learning could improve cost estimation. The new method is a machine learning model with data available in the program phase as input of the model and the construction costs as the output of the model. The output of the model is evaluated with the current estimation methods.

1.3.2 Research question

Based on the research objective the following research question is stated.

How can machine learning improve cost estimation in the program phase of the construction process?

1.3.3 Research methodology

In this thesis, the research question was worked out and structured according to Design Science Research Methodology (DSRM). Peffers et al. (2007) created this method and the process includes six steps shown in Figure 1. This method suits this research since it involves design to solve observed problems and it eventually evaluates this design. Methodological frameworks do not always tackle the problem, since these methodologies only describe a solution, resulting in a non-practical result. DSRM yields a practical design that encounters the problem. Furthermore, DSRM is consistent with previous literature.

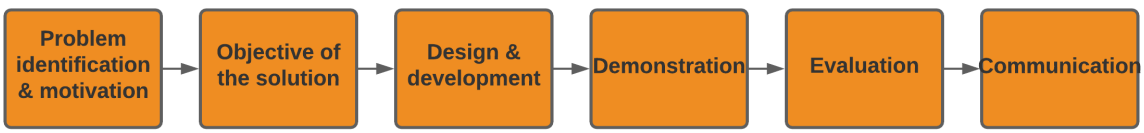


Figure 1: Design Science Research Methodology.

1.3.4 Research design

The research design consists of six stages as described in the research methodology. Stage one is the problem identification and motivation. Stage two is the definition of the objectives for a solution. Stage three is the design and development. Stage four is the demonstration, and this step is evaluated in stage five. Eventually, stage six is the communication of the results. The whole research design is visualized in Figure 2.

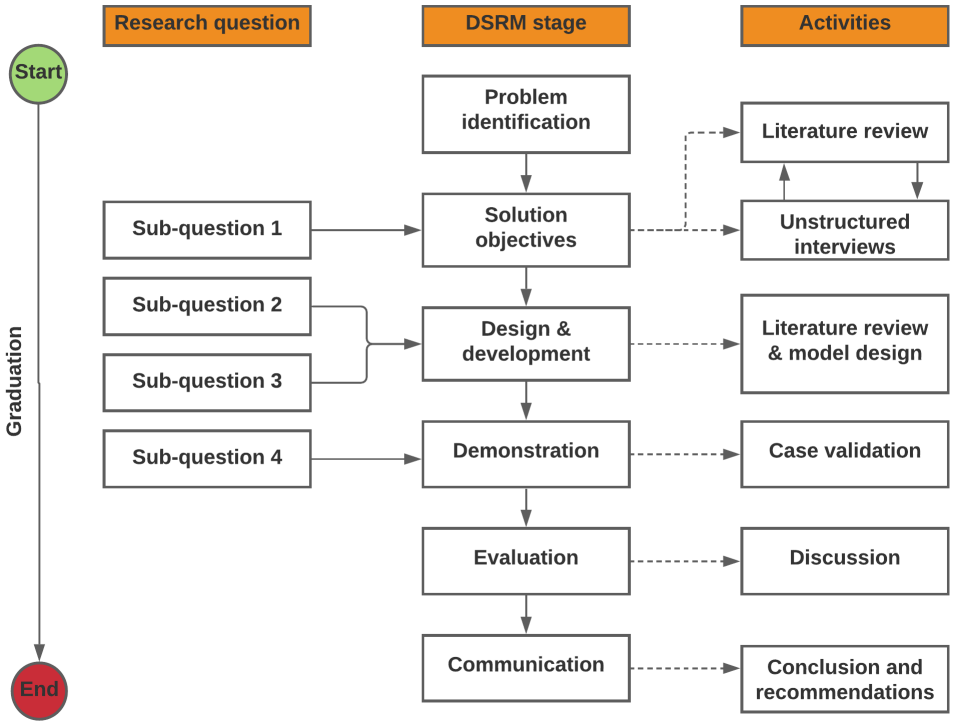


Figure 2: Research design.

Problem identification and motivation

The first step in DSRM entails defining the specific research problem and justify the value of a solution. In the research context, the research problem is derived and the potential benefits of this research are exhibited.

Definition of the objectives for a solution

This stage states the objectives for a solution based on the problem definition and knowledge of what is possible and feasible. The construction process is classified into measurable phases and the cost-related data in the program phase is observed. The conventional cost estimation method is analyzed and a norm for the accuracy of the construction costs estimation is set. This norm creates the objective for a solution. The results of the machine learning model can be compared to this norm to conclude if machine learning is a solution for the stated problem. Defining and measuring the performance of the conventional construction

costs estimation is executed by consulting literature and conducting unstructured interviews with employees of Arcadis. The objectives for this section are translated in sub-question 1 (SQ1):

1. What are the current cost estimators in the program phase in literature?
    a. What cost-related data is available during the program phase?
    b. How accurately are conventional prediction methods in the program phase?

Design and development
In this stage, the machine learning model is created. This stage consists of two parts: a literature review and the actual machine learning design. The literature review defines machine learning, investigates several machine learning types and shows the possibilities of machine learning in the construction sector. The literature review answers sub-question 2:

2. What types of machine learning can be applied to estimate construction costs?

The literature is used to construct a machine learning model that fits the data found in SQ1a. The first sections of the machine learning design concern the analyses of the data used for the machine learning design and the model development. The machine learning design phase answers the following question:

3. How can machine learning be applied to cost-related data in the program phase for cost estimation?
    a. What are the features of the cost-related data?
    b. How can machine learning process the data?

Demonstration
The last sections of the machine learning design give a demonstration and validate the results. The methodology for the validation is explained and the results from the demonstration phase are processed. This answered the following question:

4. What are the results from cost estimation using the created machine learning model?

Evaluation
Whereas the demonstration phase of DSRM validates the results, the evaluation phase interprets the results. Moreover, the evaluation discusses the practical implications of this thesis and mentions the research limitations.

Communication
In the communication phase, the conclusions are drawn. Furthermore, practical and research recommendations are given.

2 Context analysis

In order to improve construction costs estimation in the program phase, the objectives for a solution should be defined according to the second step of DSRM. The conventional method of cost estimation creates a norm, which is used as a definition for a solution. Therefore, the following main question is answered:

"What are the current cost estimators in the program phase in literature?"

In more detail, section 2.1 explains extensively how a construction process can be mapped. In addition, section 2.2 exhibits the cost-related data in the construction process, focusing on the program phase. After the construction process and its cost-related data is mapped, the general cost estimation methods are investigated in section 2.3. Section 2.4 bridges the gap between theory and practice and connects the general cost estimation methods and the actual cost estimation method. The performance of the conventional cost estimation method is analyzed in section 2.5.

2.1 Construction process

The aim of a construction process is to eventually realize a building and this is rather a complex and one-time product (Pijpers & van der Woude, 2012). The changes of failure are high when the expectations between the client and the executive are only established in the first phase of the project (Pijpers & van der Woude, 2012). Therefore, it can be convenient to structure the construction process to make it controllable during the entire process. The construction process can be approached in several ways using different methods (Pijpers & van der Woude, 2012). The method used for mapping the construction process can differ, but all have common processes. In this thesis, the method of Pijpers and van der Woude (2012) is used to describe the construction process. The method of Pijpers and van der Woude (2012) follows the norms and regulations of the *"Nederlands Normalisatie instituut"*. The data used in the demonstration section is data derived from Dutch construction projects, therefore it makes sense to use a method that is closely related to the procedure the construction projects followed when these projects were realized. However, other methods to map the construction process could be suitable as well, since methods can have common processes.

Pijpers and van der Woude (2012) divide the construction process into six main phases. Within these main phases, there can exist multiple sub-phases, that describe the main phase in more detail. The translations of all main phases and sub-phases can be found in appendix A. In chronological order, the main phases are defined as the program phase, design phase, pre-building phase, realization, maintenance and demolition. An overview of all main phases and subphases can be found in Table 2. In the following sections, these main phases will be explained. In specific the preparation phases of the construction process will be explained extensively since the focus of this thesis is to predict the construction costs. The phases after the pre-building phase focus more on the practical concerns of the construction process, which is less relevant for the scope of this thesis. The preparation phases are defined as the program phase (section 2.1.1), the design phase (2.1.2) and the pre-building phase (2.1.3), however, the remaining phases (section 2.1.4) will be mentioned shortly to complete the overview of the construction process.

Main phase	*Sub-phase*
Program	Concept phase
	Feasibility phase
	Project definition
Design phase	Sketch design
	Preliminary design
	Definitive design
Pre-building phase	Technical design
	Pricing phase
Realization	Preparation phase
	Execution phase
	Delivery phase
Maintenance	Maintenance
Demolition	Demolition

Table 2: Overview construction phases.

2.1.1 Program phase

The program phase is divided into three sub-phases: the concept phase, the feasibility phase and the project definition. The concept phase formulates the need for the construction. There can be several reasons to construct a building. However, there are also alternatives like renting or reconstructing existing buildings. Based on several factors, like financial situation, priorities and alternative solutions, a decision is made on how to meet the demands.

If the decision in the concept phase is to start building, then the next step in the construction process is to continue to the feasibility phase. In this phase, the fundamental ideas for the building are worked out. This can include the location choice, space requirement, appearance of the building, comfort and accessibility. Furthermore, the costs can be estimated very roughly, since only limited information is available.

After the fundamental ideas are elaborated, the ideas are converted to a statement of requirements. Besides this statement of requirements also the action plan is worked out. This action plan describes how the construction is eventually realized.

2.1.2 Design phase

To describe the design phase in more detail, it is divided into three sub-phases: the sketch design phase, the preliminary design phase and the definitive design phase. As said most constructions are unique, however, Pijpers and van der Woude (2012) mention explicitly that some activities are common for most buildings. For instance, designing, coordinating, planning, dividing tasks, making sketches, making calculations and completing the technical perspective. In the sketch design phase, the program of requirements is translated to an actual rough sketch. The internal and external structure of the building is determined, together with the shape and size.

After the sketch design phase, the preliminary design is created. The preliminary design defines the building spatially and functionally. This subphase of the design phase is even more detailed than the sketch design phase. Also, the exploitation costs and yield are determined.

Lastly, the design is finalized in the definitive design phase. In this phase, a very detailed design is presented. This entails the choice of material, the structure, sizes, dimensions and mechanical- and electrotechnical installations. With this design, the pre-building phase starts.

2.1.3 Pre-building phase

The pre-building phase consists out of the following two sub-phases: technical design & pricing phase. The technical design phase is focusing on the construction drawings and is less dedicated to designing. It is also the plan that is used for requesting permits. It is an accurate description of administrative tasks and the work that needs to be done. In this phase, it is possible to create a detailed cost estimation.

The main goal in the pricing phase is to find a contractor that wants to execute the plan set up in previous stages. The contractor and the client reach an agreement in this phase about the price for the execution phase.

2.1.4 Remaining phases

The remaining phases are the realization phase, maintenance phase and demolition phase. In the realization phase first, the preparations for the construction process are finished, this can concern for example logistics, but also prefabricated parts. After this, the construction process is executed and the building is rendered to the client. The building needs to be maintained during its life span and will eventually be declared unusable and demolished.

2.2 Cost-related data in the program phase

As the phases of the construction process are defined in the previous section, it is desirable to reach out more to the program phase. The program phase has limited information available since it is the first stage of the construction process. This section will specifically focus on cost-related data, the information and data available in the program phase and the type of data that could be useful for making predictions.

Throughout the stages of the construction process, different information is gathered. Bilal et al. (2016) point out that the construction sector is a sector that generates a vast amount of data. Nonetheless, not all data is useful for cost prediction, since not necessarily all information is related to the construction costs. So, it is important to select the data that is eligible for eventual cost predictions. Cost-related data is defined in this thesis as the data that has predictive value.

As described in section 2.1, the program phase consists out of three subphases. In short, these phases are the concept, feasibility phase and project definition. The focus is on the feasibility phase and project definition, since these phases concern the new building, whereas the concept phase explains the reason for building. In the feasibility phase, important activities are describing fundamental ideas, creating an investment cost estimation and exploitation cost estimation, according to Pijpers and van der Woude (2012). In the feasibility phase, only the investment costs are considerable to analyze, since these relate to the construction costs of a new building. In the project definition, the statement of requirements is fulfilled and in specific the requirements related to the object. According to Pijpers and van der Woude (2012), the object-related statement of requirements includes the organizational structure, functional units, structure and relations, norms, esthetic look and physical standards. Especially, the functional units are useful, since the building quantities can be derived from these units.

To ascertain which output in the program phase could be useful for a machine learning model, it is useful to analyze which data has been used in the construction sector to make predictions with machine learning. Chakraborty et al. (2020) analyzed the most influential factors, such as the amount of concrete used in the building to predict construction costs. Hu et al. (2021) also included factors, such as the total quantity of concrete and added more building quantities. These predictive variables, such as construction area, building height, floor height, the proportion of prefabricated components were taken into account. Huang et al. (2020) predicted the labor costs of the construction process. Huang et al. (2020) only used building quantities, such as gross floor area, basement floor area, upper floor area, first-floor area, typical floor area and effective floor area. An overview of the data used in these machine learning models can be found in Table 3.

Study	Input variables
(Chakraborty et al., 2020)	Tributary area, superimposed load, formwork & concrete.
(Hu et al., 2021)	Construction area, building height, floor height, total quantity of concrete, timber formwork rate, prefabricated components & structure type.
(Huang & Hsieh, 2020)	Gross floor area, basement floor area, upper floor area, first-floor area, typical floor area & effective floor area.

Table 3: Overview of input variables

2.3 Cost estimation methods

Cost estimation can be defined as "collecting and analyzing data in order to estimate time, money, materials and labor required to manufacture a product, construct a building or provide a service (U.S. Bureau of Labor Statistics, 2021)." More specific for the construction sector, this means cost estimators simulate the construction process and evaluate the costs of design choices (U.S. Bureau of Labor Statistics, 2021). Ali et al. (2019) describe cost estimating in more general terms as "an overview of the expected overall costs". In the context of this thesis, cost estimation will be defined as a method to forecast construction costs.

Cost estimation can be approached in different ways, to eventually forecast the construction costs as accurately as possible. Dashore (2021) appoints several methods, that can be used for cost estimation. The most relevant methods are the expert judgement method, analogous estimation method and parametric estimation method. Methods like a bottom-up approach, also known as detailed cost estimation, are not relevant for this thesis, since these methods require detailed information. This extensive information is not available in the program phase, as becomes clear in section 2.2. Each method has its strengths and weaknesses, such as accuracy, time, conditions and requirements. The different cost estimation methods will be described, so the different characteristics can be exhibited.

2.3.1 Expert judgement

Tan, Siah Yap and Yap (2012) define expert judgement as "a structured way to gather tacit knowledge to research in human cognitive behavior and communication". Expert judgement thrives on experience from previous situations, that provides valuable insights in a new, but comparable situation. This type of cost estimation is particularly useful in the program phase since only few information is available and intuition is more valuable. However, an expert opinion's vulnerability is its unstructured nature and subjectiveness (Hughes, 1996). Therefore, other more rational and argued cost estimation methods can seem more reliable and convenient. Also, if no previous comparable projects are present, this method can be less useful.

2.3.2 Analogous estimation

“Analogous cost estimating uses the values such as scope, cost, budget, and duration or measures of scale such as size, weight, and complexity from a previous, similar project as the basis for estimating the same parameter or measurement for a current project (Dashore, 2021).” Lester (2017) mentions that the major components of previous similar jobs are analyzed and included when estimating the costs of a new, but comparable project. This way analogous cost estimations can become more accurate. Dashore (2021) points out this way of cost estimation is particularly useful in combination with the right expertise. Drawbacks from this method are lack of comparable constructions or not enough expertise to translate these comparable constructions into a new construction cost estimation. Also, according to Matel (2019), this way of cost estimation has limited accuracy.

2.3.3 Parametric estimation

Parametric cost estimation uses algorithms or statistics to determine a relationship between variables, based on historical data (Dashore, 2021). This relationship can be used to estimate costs when some variables are already known, but others are not. Parametric approaches make use of regression, factor- and principal component analysis to elicit the important parameters that influence construction costs (Swei et al., 2017). The most desirable situation is when there is a strong relationship between variables. This way it is possible to accurately determine one variable if only the other variable or variables are known. At the same time, this can be a limitation using parametric estimation, since the absence of a strong relation, most likely leads to an inaccurate estimation. Moreover, it is important that this relation is not taken for granted, since it can change throughout the years. When variables fluctuate, it can be necessary to adapt when performing a parametric estimation, for example when building materials get more expensive throughout the years. Furthermore, (Matel, 2019) points out it can be difficult to perform correct statistical techniques that are needed to build a quality model.

2.4 Conventional construction costs estimation

Section 2.3 describes generally the types of estimation; this section elaborates further on how construction cost estimation methods are practiced. To close the gap between literature and practice, this section makes use of a combination of literature and unstructured interviews with cost specialists from Arcadis. The unstructured interviews were conducted in Dutch. Afterwards, the interviews were translated and summarized. The unstructured interviews can be found in appendix B.

The U.S. Bureau of Labor Statistics (2021) mentions construction cost estimation is often executed using databases and other records to compare the costs of similar projects. The interviews that were conducted confirm this. One respondent mentions that the type of building is determined and referred to other similar projects. Also, the same respondent concludes that variables of some type of buildings will stay within a certain bandwidth, which makes it particularly suitable for cost estimation. Multiple respondents point out they use cost key figures to estimate the construction costs. Key cost figures can be defined as numbers that help to estimate construction costs, based on historical data and actual prices. With the building properties, that can be derived from the statement of requirements, and the key cost figures combined it is possible to estimate the construction costs.

Based on literature and unstructured interviews, it is most convenient to conclude that the conventional cost estimation method is a combination of analogous estimation and expert judgment. The current cost estimation accommodates itself to the characteristics of an analogous estimation since historical data is used and important components of earlier buildings are analyzed. However, as mentioned, in absence of an

expert the analogous cost estimation is less effective since specific knowledge is required. A machine learning model can be classified as a parametric estimation since machine learning uses statistics and historical data to estimate construction costs. An improvement of cost estimation, due to machine learning, may cause a shift from analogous/expert estimation towards parametric estimation.

2.5 Performance of cost estimation methods

As can be concluded from section 2.4, conventional construction costs estimation complies mostly with a combination of analogous- and expert judgement estimation. This section will focus on the performance of the conventional cost estimation method. The performance is measured in accuracy and the time used to apply the cost estimation method. Unstructured interviews (appendix B) are used in combination with literature to report on the performance of the conventional method.

Lester (2017) argues that in the feasibility phase the percentage error of conventional cost estimation can reach 10 to 20% and even mentions that the percentage error can reach between 5 and 10% in the project definition phase. However, cost specialists all argue the percentage error in the project definition phase is an optimistic number. The respondents all gave an answer in the range of 10-20% for more standard and simple buildings, like for example offices and houses. According to multiple respondents, the error range depends on the complexity and nature of the project. One respondent points out that for more unique and special buildings the percentage error can lay between 20-30%. Moreover, a respondent explains that the accuracy of the estimation depends on the information available thus far. Project Management Institute (2016) supports this respondent and mentions that the accuracy depends on how well the scope of a project is defined.

The time respondents spend on cost estimations is variating. One respondent remarks that a cost estimation for a type of building of his specialty only costs one hour. On the other side, multiple respondents comment that a complex project can take far more time, even up to three days. According to a specialist, the most time-intensive process is in extracting information out of the statement of requirements. Also, the availability of previous projects is important. If the information is retrieved out of the statement of requirements it is compared with earlier projects. The absence of comparable projects leads to spending a longer time on the cost estimation.

2.6 Conclusion

The construction process can be divided into six main stages, which again can be divided into more substages. In the program phase, there are several information streams, however, not all information is suitable for cost estimation, since not all information is related to the construction costs. In the program phase, the statement of requirements is especially interesting to extract cost-related data from, such as building quantities. Conventional construction costs estimations also use this statement of requirements in combination with an analogous estimation and expert judgement to assess the construction costs. The performance of the conventional method differs depending on the type of project and is not uniform for all projects. Factors that are influencing this percentage are for example the nature of the project and how well the scope is defined. Standard and simple buildings can be estimated with a 10-20% error, whereas more complex buildings can be estimated with a 20-30% error. Following DSRM, these error ranges define the objective for a solution and can be used as a norm to compare the created machine learning model with. The time spend on a cost estimation varies from hours to multiple days, depending on the expertise and complexity of the project. However, the most time-costly process is extracting information from the statement of requirements.

3 Literature review

The previous section defined an objective for a solution. This section elaborates on the literature that is needed for the design & development, stage three of DSRM. This literature study has been executed to gain more information about machine learning and in specific the types of machine learning that are relevant for cost prediction. The several methods were discussed, compared and connected to the construction sector. The following question is discussed in this section:

"What types of machine learning can be applied to estimate construction costs?"

Section 3.1 defines and explains machine learning in general and describes model validation techniques. Section 3.2 describes the types of machine learning applicable to cost prediction. Section 3.3 bridges the gap to machine learning in the construction sector.

3.1 Machine learning

3.1.1 Definition

Machine learning (ML) is a subdiscipline of Artificial Intelligence (AI), which attempts to emulate how a human brain understands and interacts with the world (Lawson et al., 2021; Maheshwera et al., 2020; Osarogiagbon et al., 2021). This means theoretically that a fully functioning AI can perform the same as the human brain. In Lawson's context features of AI can be self-driving cars, recommending medical treatments or summarizing texts like humans. Speaking about machine learning we always talk about AI, but this is not the case vice versa, as can be seen in Equation 1.

Machine learning is in the literature acknowledged as the study of computer algorithms that seek to improve automatically through experience (Lawson et al., 2021). Machine learning can determine patterns and correlations and discover knowledge from datasets using historical data during the training phase (van Klompenburg et al., 2020). Machine learning can even identify patterns invisible to human eyes (Liu et al., 2019). Overall, this makes machine learning a powerful tool for processing datasets and recognize patterns. Machine learning has the peculiarity of being able to make predictions even if it is not specifically programmed to perform the task, because of the mathematical model that has been trained with data (Mazzi, 2021). The training of the machine learning model is done with the majority of the historical data available; this is what is acknowledged as "training data" or "training sets". The leftover historical data is the "test data" or the "test sets" (Fenza et al., 2021). This "test data", that the machine learning has not seen before, can be used to indicate the accuracy of the machine learning model (Mangalathu et al., 2021).

In literature, neural networking and deep learning are often individually addressed, without mentioning they are a form of machine learning. Neural networks and deep learning models are often used to reconstruct complex relations between quantities selected as input and output (Pasini, 2021).

$$\{Deep\ learning\} \in \{Neural\ Networking\} \in \{Machine\ learning\} \in \{Artificial\ intelligence\}$$

Equation 1: Subsets of Artificial intelligence, source: (Lawson et al., 2021).

3.1.2 Approaches

Machine learning approaches can be divided in three subcategories, depending on the purpose of the model. These three approaches are classified as supervised learning, unsupervised learning and semi-supervised learning. The difference between the different approaches depends on the signals or feedback that the models get. Supervised methods are labelled examples and then used to make predictions, whereas unsupervised learning structure in a dataset without using labels (Bao et al., 2019). The semi-supervised learning method requires labels but also makes use of unlabelled examples (Libbrecht & Noble, 2015). Ghatak (2017) explains this in more detail, as that the algorithm of unsupervised learning gathers experience from a data set that contains only features, without any dependent variable. Supervised learning does not only contain features, but also a target variable, which is a function of the feature. Both supervised and unsupervised learning have their own applications. The objective of supervised learning is to estimate the target variable of the model, such that the output error is minimized, while the objective of unsupervised learning is to build representations of the inputs that can be used for decision making and predicting future inputs (Suresh et al., 2013).

Within supervised learning and unsupervised learning, we can also appoint the algorithms as continuous or discrete (Urbanowicz & Browne, 2017). The fundamental difference is that within a continuous model, the output is an element of all real values. The discrete model has a discrete-valued output, where the output is placed in a category. Combining supervised and unsupervised learning with the continuous or discrete characteristics results in the following Table 4.

	Unsupervised	**Supervised**
Continuous	Dimensionality reduction	Regression
Discrete	Clustering	Classification

Table 4: Classification of machine learning approaches, based on (Reel et al. 2021).

In the case of this thesis, we want to predict the construction costs in the program phase, Table 5 gives a demonstration of data available in the program phase. Since construction costs are a real value, this results in a continuous machine learning model. Furthermore, “construction costs” is the target variable of the dataset and the dataset includes historical constructions costs that can be predicted in a machine learning model, therefore the model can be classified as a supervised machine learning model. Combining these two properties results in a regression machine learning model, as follows from Table 4.

Project	**Gross Floor Area (GFA)**	**Gross Building Volume (GBV)**	**External Wall Openings (EWO)**
1	3280	11127	946
2	14533	42456	4666
3	1182	4224	320

Table 5: Example data.

There exist more approaches such as reinforcement learning. However, Azlinah (2020) describes that most articles recognize only supervised and unsupervised learning since other algorithms such as reinforcement learning also take into account the environment. These approaches are also not eligible for this thesis.

3.1.3 Model validation

Whereas the previous sections define machine learning, this section focuses on the validation of the model. To assess if the machine learning model is generalizable it is important to validate the model. Section 3.1.3.1 explains several methods to measure the error using loss functions. Section 3.1.3.2 explains the statistical validation named "cross-validation" and section 3.1.3.3 elaborates on under- and overfitting.

3.1.3.1 Loss functions

A model is only useful if it reflects the actual situation. The goal of a machine learning model is to predict the situation right. However, if a machine learning model does not fit the situation, most likely the prediction is also deviating from the actual situation. As Stetco et al. (2018) mention, many machine learning models reported in literature lack proper validation procedures. Recapitulating section 3.1.1, to evaluate a model the data should be split into a "training set" and a "test set". When the machine learning model is properly trained by the training set, it is possible to evaluate by using the test set. If the model would reflect the relationship between variables, the test set would show a minimum loss. The validity of regression-based machine learning models is based on the predicted output and the actual output (Stetco et al., 2018).

The difference between the machine learning models' output and the actual output can be described by a loss function (Maleki et al., 2020). Logically, the loss function should be as small as possible, since the gap between the predicted value ($\hat{y}$) and the actual value (y) should be as close as possible. Regression loss functions and other loss functions, such as classification loss functions, should not be confused. Classification loss functions focus on discrete values and regression loss functions on continuous values. Brownlee (2019) mentions three regression loss functions. These three loss functions are the Mean Squared Error (MSE), Mean Squared Logarithmic Error (MSLE) and Mean Absolute Error (MAE). All these different loss functions measure the "error" in a different way, resulting in a different outcome. These three loss functions will be discussed further in this section.

(Root) Mean Squared Error ((R)MSE)

Many machine learning algorithms are based on the Mean Squared Error (Vandeput, 2019). The MSE function measures the squares of error with every iteration. The MSE is always positive since the error is squared. Just like almost any other loss function, the closer this function heads to zero, the more accurate the predicting of the machine learning model is. However, this loss function is less suitable when there are many outliers since an outlier creates a big difference between the predicted value and the actual value. Using the MSE, this difference is squared, and the error can become high. The function for the MSE is shown in Equation 2.

$$\boldsymbol{MSE} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}(\boldsymbol{y_i} - \hat{\mathrm{y}}_i)^{\mathbf{2}}$$

Equation 2: MSE

However, as can be derived from the MSE, the error resulting from the MSE formula is squared. It can be overwhelming to interpret if the average error is high. Therefore, an extension of the MSE method is the Root Mean Squared Error (RMSE). This formula scales the error and can be denoted as shown in Equation 3.

$$\boldsymbol{RMSE} = \sqrt{\boldsymbol{MSE}}$$

Equation 3: RMSE

Mean Squared Logarithmic Error (MSLE)

The Mean Squared Logarithmic Error (MSLE) is the logarithmic value of the average of the difference between the predicted values and the actual values. This type of prediction is especially useful in areas with big outliers as well as in areas with small outliers and calculates an error by assigning penalties to items with underestimated outliers rather than overestimated outliers (Jeong et al., 2020). The equation of the MSLE is shown in Equation 4.

$$\boldsymbol{MSLE = \frac{1}{n}\sum_{i=1}^{n}(log(\hat{y}_i + 1) - \log(y_i + 1))^2}$$

Equation 4: MSLE

Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) shows the average of the absolute errors between the predicted values and the actual values (Jeong et al., 2020). The MAE gives the same result as the RMSE. Comparing the MAE to the MSE, the MAE is more relent to outliers, since it does not square the error, but only makes it an absolute value. Its formula can be seen in Equation 5.

$$\boldsymbol{MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|}$$

Equation 5: MAE

In addition, an often-used loss function that acts the same is the Mean Absolute Percentage Error (MAPE). However, the difference is that the MAPE compares the error against the actual value, which can make it a useful function when the data is more deviated. Also, the MAPE is denoted in percentages, which makes it better interpretable. The MAPE can be noted as can be seen in Equation 6.

$$\boldsymbol{MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|}$$

Equation 6: MAPE

3.1.3.2 Cross-validation

To prevent that in the validation phase the error gets biased, cross-validation can be used. Lu et al. (2019) describe cross-validation in a machine learning context as evaluating the accuracy of the model by using testing data to make predictions with the already trained machine learning model. There are multiple ways of using cross-validation, each having its own advantages (Lu et al., 2019). Instances of cross-validation methods are K-fold, repeated random subsampling and hold-out method (Dubitzky et al., 2007). Using K-fold, the dataset is split into a certain number of minor datasets, this number of datasets is the value k. The value k depends on the balance between the test set and the training set, since the preference is to maximize the amount of data both in the test and training set, the balance can differ. After this, each time a part of the data is used as test data and the remaining part as training data. A limitation is that only the exact number of data available can be used. Repeated random subsampling does not have this problem. It randomly

selects training and testing data, resulting in that an unlimited of iterations can be performed. However, randomly selecting data could potentially mean that some data points can be selected more often than others and some points can be selected not at all. These are two relative complex methods, whereas the hold-out method is less complex. It only concerns one training and test set, resulting in a simple type of cross-validation. However, this simplicity also results in a highly biased method (Dubitzky et al., 2007).

3.1.3.3 Underfitting and overfitting

Overfitting and underfitting can be problematic for machine learning models. Figure 3 visualizes the problem of underfitting and overfitting. An underfit machine learning model follows the data poorly and therefore does not represent the relation between the variables in a dataset. When the machine learning model does not comprehend the training data, nor will it with the test data (Czajkowski & Kretowski, 2019). The opposite of underfitting is overfitting. Whereas underfitting has a lot of variation, overfitting reduces the variation. This results in a model that is very specific and not generalizable for the whole dataset (Ellingson et al., 2020). The best situation is the "good fit" situation. This takes into account variation by finding the right balance, this way the model is generalizable for a new data point, as can be seen in Figure 3. There are several measures to check overfitting and underfitting. An example of a measure is cross-validation (Javatpoint, 2021).

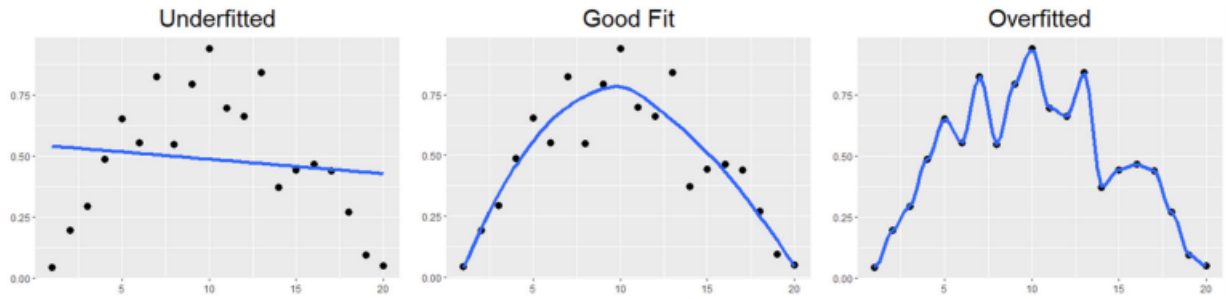


Figure 3: Underfit, good fit and overfit, source: (Nanda, Vallmuur, and Lehto 2018).

3.2 Types of machine learning

Machine learning is not one type, on the contrary, there exist nowadays multiple machine learning methods in the public domain (Lawson et al., 2021). However, every machine learning model has its own characteristics and no single type of machine learning is best for every learning task (Lawson et al., 2021). Some machine learning algorithms may be more robust than others in terms of accuracy, but there are other factors that are important to consider for selecting a machine learning method. Factors like the implementation speed, ease of interpreting results, nature of supervised machine learning task, nature and amount of input and output data and the complexity of the model to be learnt are crucial factors to consider when deciding on the machine learning algorithm (Osarogiagbon et al., 2021). This section will focus on the supervised machine learning methods with a real value as output, also familiar as a regression machine learning model.

In some regression models, the value of the output is based on probability (Eyo & Abbey, 2021). To explain this in more detail, unlike regression methods, the output is classified. This means that based on the input, the model will choose a certain category and depending on the estimation method of that category, the output will get a certain value. As may become clear, these models seem to apply more to the classification quadrant than to the regression quadrant, explained in section 3.1.2. However, Chakraborty et al. (2020) point out machine learning types, such as decision trees, are the most contemporary machine learning types for construction costs estimation. While decision trees can give a discrete or continuous output (Shakil et al., 2021).

Maleki et al. (2020) describe fundamental and classic machine learning approaches, which are reviewed further in this thesis. The supervised machine learning types, that Maleki et al. (2020) describes, are linear regression, support vector machines, K-nearest neighbor, decision trees, random forest, naïve bayes and neural networks. Understanding the methodology of the different regression machine learning models and other characteristics will help to eventually decide on what machine learning model is most suitable for this thesis.

3.2.1 Linear regression

Linear regression is one of the most widely and intensively used techniques of quantitative research in any area of research where one is interested in studying the relationship between a variable of interest, called the response variable and an element of predictor variables (Gujarati, 2019). Equation 7 shows the linear regression model, where b_0, b_1, b_2, ..., b_n equals the coefficient(s) of the line, $h_b(x)$ is the predicted value and x_1,x_2, ... , x_n is the actual value of the predictor(s). However, assuming x_0 equals one, it can be convenient to write it down in matrix form as can be seen in Equation 8.

$$h_b(x) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n$$

Equation 7: Linear equation

$$h_b(x) = [b_0 \quad b_1 \quad \dots \quad b_n] \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} = b^T x$$

Equation 8: Linear equation in matrix form

The error of the linear regression model is the distance between the linear line and the actual values, as can be seen in Figure 4. There are several methods to minimize this error. Scikit-Learn, a library with a wide range of supervised machine learning algorithms, offers multiple estimators that minimize the error. Ghatak (2017) uses the least ordinary squares, with as a result that outliers are causing heavy errors, since the error is squared. RANSAC does not have this problem since the model is only estimated from the determined inliers. (Scikit Learn, 2021). It can be concluded that the estimator choice depends on the data.

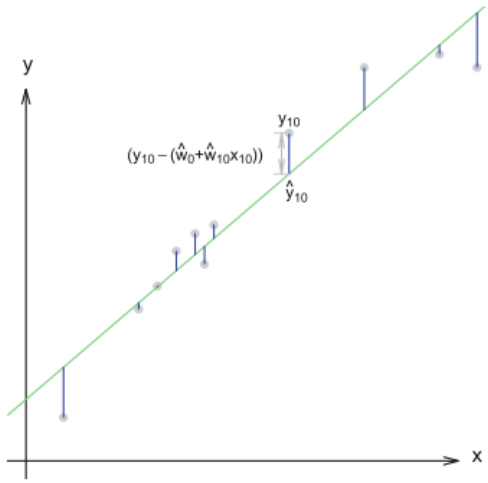


Figure 4: Minimizing the error. The green line is the linear regression line and the blue lines are the difference between the actual values and the predicted values, source: (Ghatak, 2017).

Ghatak (2017) points out that there are a few important assumptions that need to be made for applying the linear regression method. The first one and directly the most logical one is that the predictor variable and predicted variable should have a linear relationship. The absence of a linear relationship leads to a misfit of the regression line and the actual points. The second assumption is the absence of collinearity within the independent variables. This is the case when two variables are dependent on each other. This is not necessarily a problem but makes it difficult to report the individual effects of the predictor variables.

3.2.2 Support vector machines

Support vector machines (SVM) is a machine learning technique that uses hyperplanes to separate the data points in space. These hyperplanes are the boundaries of the machine learning model and thus classifies the different data points. The lines of the hyperplanes are linear, however, there are tricks in SVM that also deal with data that is not linearly separable. Even though SVM is a classification algorithm, Maleki et al. (2020) explicitly mention that SVM can be adopted and used for regression problems.

3.2.3 K-nearest neighbor

K-nearest neighbor classifies a new unlabeled data point, based on the K labelled data points that are the most similar to the new observation (Maleki et al., 2020). K-nearest neighbor (K-NN) is commonly used as classification machine learning, although it can be used for regression as well (Maleki et al., 2020). Logically, a disadvantage of K-NN is that it is rather a classification algorithm than a regression algorithm. Thus, the output depends on the estimation of the class the new value is assigned to.

3.2.4 Decision trees

The name of the decision tree resembles what the machine learning algorithm is. Each node of the decision tree represents a point to make a decision and based on the outcome it will choose a branch. At the first

node, all data has to be divided into branches and later in the process, the data points will be appointed to the best class. This best class is in the decision tree algorithm defined as the leaf node and is the node where the sample meets a stopping criterion. Figure 5 visualizes this process. Maleki et al. (2020) describe that in regression problems often two new nodes are generated and the samples with a value less or equal to the threshold are assigned to one node and the rest of the samples are assigned to the other. An advantage is that pre-processing is not necessary for decision trees, but a disadvantage is that decision trees achieve a poor performance for imbalanced datasets and can easily overfit (Maleki et al., 2020).

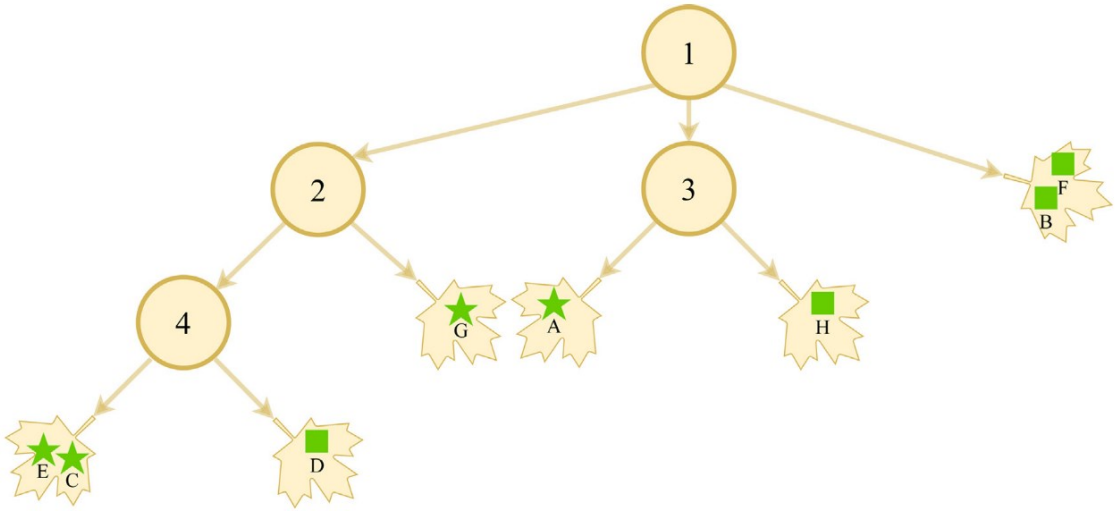


Figure 5: Decision tree, source: (Maleki et al., 2020)

3.2.4.1 Random forest

Random forest consists of a collection of decision trees. The added value of a random forest compared to decision trees is that random forest can deal with high variance (Maleki et al., 2020). Each tree is independently trained on random subsamples of the training set. Random forest results in fewer correlations across the decision tree (Maleki et al., 2020). This improves the performance of the resulting machine learning model.

3.2.5 Naïve bayes

Naïve bayes is a probabilistic machine learning approach. *"In order to predict the class of unknown observations, these approaches rely on the naive assumptions that each pair of features are conditionally independent given their corresponding class label, which means that, for a given class, each feature is statistically independent of the other features; that is, the value of a feature x_i is not related to the value of a feature x_j. Naïve.* (Maleki et al., 2020)*"* However, naïve bayes is a generative model, which means it learns from the joint probability to calculate the conditional probability P(A|B), while other machine learning models estimate this directly from the training data (Nanda et al., 2018).

3.2.6 Neural networks

Kaviani and Sohn (2021) describe typical neural networks as complex networks which consist of a large number of neurons that are interconnected together by weighted links. The neurons of different layers in the network are connected and depending on sufficient input a neuron will activate and send an output to the next layer in the neural network. Logically, the weight of a connection increases when the signal is often passed to the next layer. The first layer is classically defined as the "input layer" or "predictor variables" as

we refer most of the time to in this thesis. The last layer is the “output layer” or “predicted variable”. In between these layers, there can be many hidden layers, that are trained to come to an output. Figure 6 illustrates how a typical neural network can be build up, however, the number of layers and neurons can vary.

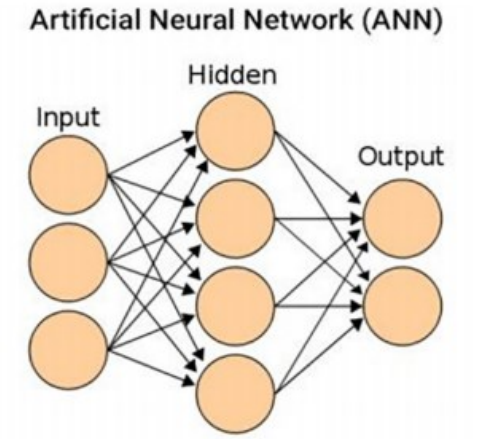


Figure 6: Example of an ANN, source: (Desai & Shah, 2020).

How powerful these networks may seem, they bring also serious drawbacks. The learning performance of neural networks depends on the structure and the learning algorithm applied to the network (Kaviani & Sohn, 2021). Moreover, Lawson et al. (2021) explain that neural networks do not perform better than simple machine learning algorithms when less than 1000 labelled data instances are present, as can be seen in Figure 7. Besides, the computation power needed for neural networks can be insufficient as more data is added (Lawson et al., 2021). Also, Jawad, Hawari and Javaid Zaidi (2021) point out that there is no standard way to determine the network architecture and mostly trial-and-error is used. Furthermore, an improperly trained network may converge to a local minimum (Jawad et al., 2021). In short, there are numerous drawbacks when selecting a neural network. It is important to realize these consequences when choosing a neural network as machine learning type.

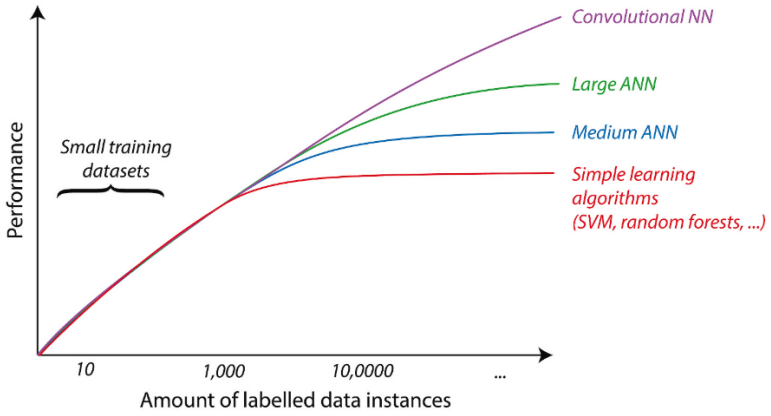


Figure 7: Data drives deep learning performance (Lawson et al., 2021).

3.3 Machine learning in the construction sector

Machine learning is already used in the construction sector for several purposes, like thermal design optimization (Yigit, 2021), bridge inspection (Dorafshan & Azari, 2020) and safety issues (Choi et al., 2019). This section scopes down to the use of machine learning in the construction sector. More specifically to cost estimation and its performance in the construction sector.

Kim et al. (2004) predicted the construction costs using linear regression and neural networks. The input data like gross floor area, stories, total unit, duration and roof types. Again, the neural network outperformed the linear regression model, as the MAPE of the linear regression model was 6.95% and MAPE of the neural network was 2.97% in its best model. According to Kim et al. (2004), a limitation of the neural network model was the finetuning, which was a time-consuming process. Without this finetuning, the neural network model had a MAPE comparable with the linear regression model.

Alshamrani (2017) predicted the construction costs with a linear regression model. The input parameters considered were building area, floor height, number of floors and structure & envelope types. The Average Invalidity Percentage (AIP) found was 5.6%. In this study, not many outliers in the percentage error were detected, which means that the model had a good consistency on the test data. Furthermore, the linear regression model was supported by normalization, which caused outliers to have less influence on the machine learning model.

Elfahham (2019) predicted Construction Cost Index (CCI) using machine learning techniques, like neural networks and linear regression. CCI consisted out of four building materials, each having its own weight. These weights were calculated by determining the quantities of the four material components based on their product values and unit prices during the base year. Time series (a series of data points in time), neural networks and linear regression were used to predict the test set. Time series performed best as it had a MAPE of 3.5%, whereas the neural network and linear model had a MAPE of 8.3% and 17.5% respectively. However, an important side note is that the CCI had a stable trend except for the last three years, which showed an unusual jump due to government regulations (Elfahham, 2019).

Overall, the results of machine learning in the construction sector are promising, as can be seen in Table 6. However, several cases of construction cost prediction confirm that every machine learning model has its own characteristics and no single type of machine learning is best for every learning task, as pointed out by Lawson (2021). Also, the statement of Osarogiagbon et al. (2021) is in line with what the cases state. Multiple criteria must be taken into account when selecting a machine learning method.

Authors	Model	Error
(Kim et al., 2004)	Linear regression & Neural networks	MAPE: 2.97% (Neural networks, best model) MAPE: 6.95% (Linear regression)
(Alshamrani, 2017)	Linear regression	AIP: 5.6% (Linear regression)
(Elfahham, 2019)	Time series, Neural networks & Linear regression	MAPE: 3.5% (Time series) MAPE: 8.3% (Neural networks) MAPE: 17.5% (Linear regression)

Table 6: Overview of machine learning cost estimation

3.4 Conclusion

Machine learning is in literature acknowledged as the study of computer algorithms that seek to improve automatically through experience. Machine learning has the peculiarity of being able to make predictions even if it is not specifically programmed to perform the task, because of the mathematical model that has been trained with data. Supervised machine learning has labelled data, whereas unsupervised machine learning has unlabelled data. An output variable can be continuous or discrete. Since construction cost is a target variable and continuous value, it can be concluded that construction costs prediction has a regression machine learning approach. There are several methods to validate the model to test if it is generalizable. The difference between the machine learning models' output and the actual output can be described by a loss function. To prevent that in the validation phase the accuracy gets biased, cross-validation can be used. Also, if a model is underfitted or overfitted, it is biased and not generalizable since it does not deal well with variance (underfit) or does deal with variance too much (overfit). Cross-validation can be used to evaluate this.

There exist nowadays a lot of machine learning types in the public domain and every machine learning type has its own characteristics. Besides accuracy, there are several other factors that should be taken into account when selecting a machine learning type. Some machine learning types are more defined as classification models, but practice shows that the models are also suitable for regression problems.

Earlier studies in the construction sector that used machine learning offer promising results regarding the accuracy of the models. However, the context of these studies reveals the importance of multiple criteria when selecting a machine learning type. Not one single machine learning model is the best for every learning task and each machine learning model has its advantages and disadvantages.

4 Machine learning design

Thus far, in DSRM the problem is identified & motivated, the objective of the solution is defined and a literature review that is needed for the design & development stage is performed. This section includes the next steps of DSRM: the actual design & development of the machine learning model and the demonstration of the model. In order to be able to compare different estimation methods, the model should be created and the results should be analyzed. Therefore, the following questions are answered:

"How can machine learning be applied to cost-related data in the program phase for cost estimation?"

"What are the results from cost estimation using the created machine learning model?"

Before it is possible to select a suitable machine learning model, it is necessary to have an appropriate data analysis. In Section 4.1, the data will be explained and analyzed. In Section 4.2, the actual model is selected and developed. The results of the developed model are shown in Section 4.3.

4.1 Data analysis

To know what model to choose eventually, there should be an appropriate data analysis first. This section describes which predictors are used (section 4.1.1) to eventually predict the construction costs. These construction costs are until now explained very generally, however, section 4.1.2 will focus more on specific construction costs. Section 4.1.3. broadly approaches the data and shows its characteristics, whereas section 4.1.4 uses the previous section to prepare the data for the machine learning model.

The data used to create a machine learning model is gathered from Arcadis and is extracted from Dutch construction projects from 1997 until 2017. Section 2.3 elaborates the fact that variables can fluctuate throughout the years. Therefore, the data used is indexed, resulting in that the data can be compared over several years. The data used in this thesis is indexed structural and cyclical. Structural indexing compensates structural raise of construction costs, whereas cyclical takes into account the time-specific raise of construction costs. In total 1019 projects are included in this thesis.

4.1.1 Predicting variables

From previous literature focused on the construction sector (section 2.2), we can conclude that several predicting variables could be useful for making machine learning predictions. Reflecting on the program phase, not all predictors are suitable for this phase. For example, the type of material used can be included in the statement of requirements, nevertheless, it is difficult to determine the volume of materials. However, building quantities such as gross floor area and building height can be derived from the statement of requirements. Besides, the literature suggests that these building quantities could potentially work out as predictors. Therefore, the building quantities in the program phase are used further in this thesis.

In the dataset used for this machine learning model, nine building quantities are present, which can possibly have a correlation between the total construction costs. These key costs figures will be the predicting variables in the machine learning model. The predicting variables can be seen in Table 7, whereas the last column represents the percentage of the 1019 projects where the particular variable is registered. The relation between the predicting variables and the construction costs will be analyzed in section 4.1.3.

Number	Predicting variables	Abbreviation	Availability in projects
1	Gross Floor Area	GFA	99.51%
2	Gross Building Volume	GBV	91.81%
3	Built-up Area	BUA	97.08%
4	External Walls Gross Area	EWGA	97.41%
5	External Wall Openings	EWO	95.52%
6	Gross Roof Area	GRA	92.79%
7	Enclosed Area	EA	93.96%
8	Internal Walls Gross Area	IWGA	90.74%
9	Internal Wall Openings	IWO	89.86%

Table 7: Predicting variables.

4.1.2 Construction costs

Until now, always the term *"construction costs"* has been used in this thesis. To go into further detail, the construction costs in the dataset can be divided into several sub costs, that create together the total construction costs. It is relevant to define which specific costs components are included when the term *"construction costs"* is used in this thesis since this creates transparency about what is actually predicted. An overview of the components that are included in the total construction costs can be found in Table 8. In this thesis, all components of the total construction costs, mentioned in Table 8, are included for predicting the construction costs.

Number	Component of total construction costs
1	Demolition and site preparation Costs
2	Substructure Costs
3	Structure Costs
4	External Walls Costs
5	Roofs Costs
6	Fitting out/Built-in Costs
7	Mechanical Installations Costs
8	Electrical Installations Costs
9	Transport Services Costs
10	Equipment Costs
11	External and ancillary works Costs
12	Overheads Site Costs
13	Special Conditions Costs

Table 8: Components of construction costs.

4.1.3 Pre-processing

Out of the 1019 projects that are available in the dataset, it is necessary to divide the projects into categories. Projects can have a different origin, which also results in other correlations between the project and the construction costs. Therefore, the projects can be divided into basically two categories: residential projects and non-residential projects. 582 projects are defined as residential projects and the remaining 437 projects are defined as non-residential projects. Within these categories, it is convenient to create classes since different building classifications are not comparable and will result in a less accurate and reliable model. The residential projects in the dataset can be divided into 5 classes, as can be seen in Table 9. However, non-residential buildings are far more complex and can be divided in the dataset into 56 classes. Examples of non-residential buildings are educational buildings and offices. The combination of fewer projects available in total in the non-residential category and the high number of classes results in a limited amount of data per classification in the non-residential category. More data leads to a better model performance (Lawson et al., 2021). Therefore, in this thesis, the scope is on the residential buildings to improve the accuracy of the predictions.

Overall, the number of projects in the residential category is 582. Table 9 elaborates more specific on the number of projects per class available in de dataset. Apartment blocks with an internal entrance and terraced houses represent together many projects, whereas the other three classes do not. More data leads to a more accurate prediction. Therefore, it is convenient to focus on the two classes with most of the data. The data of apartment blocks with an internal entrance and terraced houses is used for the machine learning model. More precise, this means that the building quantities of apartment blocks with an internal entrance and terraced houses are used as input to predict the construction costs of these two separate building classes. In short, this means that there will be two data sets, two separate trained machine learning models and two outcomes. Eventually, it is possible to compare the machine learning results of apartment blocks with an internal entrance and terraced houses to see if one class is better predictable. The data will be prepared in section 4.1.4.

Number	Classification	Number of projects	Percentage of total
1	Apartment Block (internal entrance)	261	44.85%
2	Apartment Block (gallery entrance)	25	4.29%
3	Terraced House	214	36.77%
4	Detached House	62	10.65%
5	Semidetached House	20	3.44%
Total		582	100%

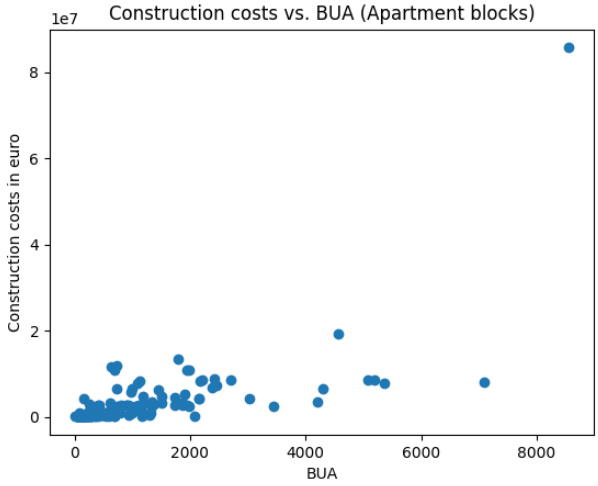
Table 9: Classification residential constructions.

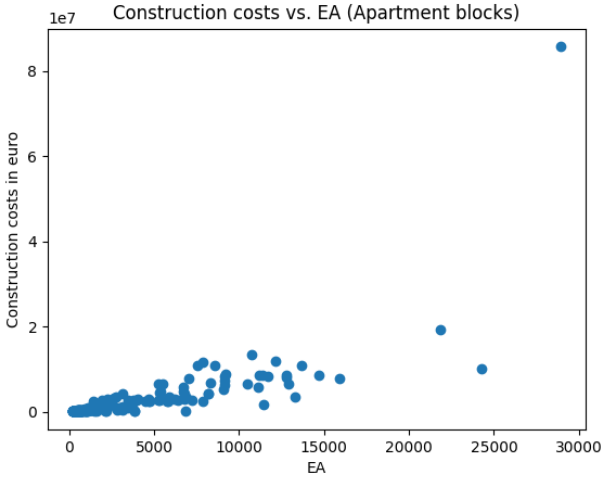
In order to understand the cost characteristics of the classes that have been selected for the machine learning model, an analysis is performed. Table 10 shows descriptive statistics of the construction costs of the two selected classes. These statistics include the maximum, minimum, median, mean and the standard deviation. In both classes the difference between the minimum and maximum is high. However, noticing that the median is between the mean and minimum, most of the data is closer to the minimum. Therefore, there will be a couple of outliers in both classes.

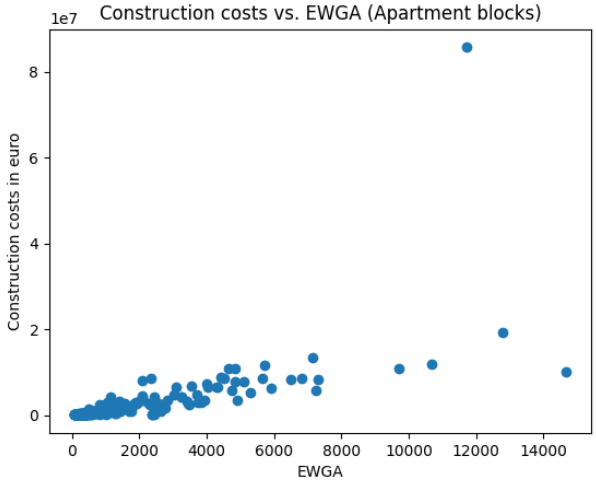
Statistic variable	Apartment Block (internal entrance)	Terraced House
Maximum	€ 85.656.808	€103.292.635
Minimum	€50.050	€50.700
Median	€312.100	€786.700
Mean	€2.554.736	€5.015.448
Standard deviation	€6.872.252	€10.440.406

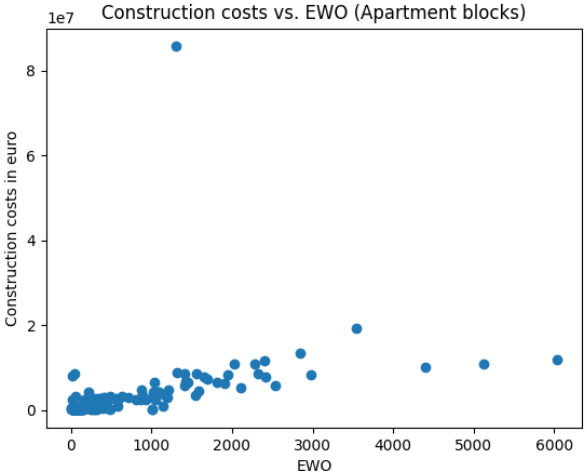
Table 10: Summary statistics apartment blocks & terraced houses.

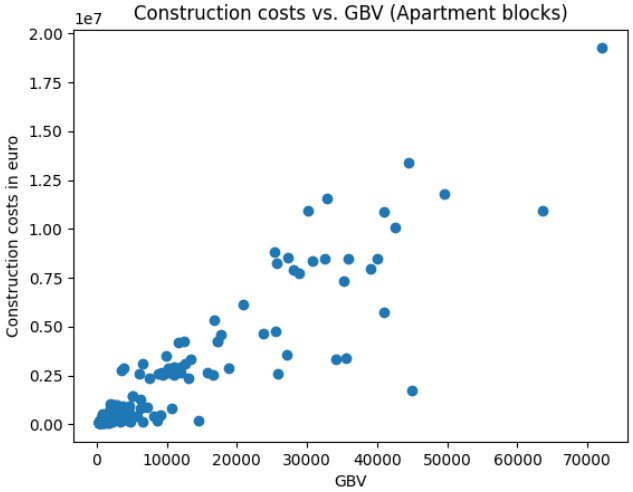
To determine if there is a correlation between the predictors and the construction costs scatter plots are used. If a correlation is present, these scatterplots are useful to visualize patterns in the data. The scatter plots that are shown below originate from apartment blocks with an internal entrance. The scatter plots of the terraced houses are shown in appendix C. On the y-axis the total construction costs can be found, for almost all building quantities the y-axis is scaled the same. On the x-axis, the different building quantities are represented. The scatter plots clearly reveal a linear pattern for the apartment blocks and terraced houses and some building quantities. Other kinds of patterns are not detected in the data. Some building quantities have a more linear relationship with the construction costs than other building quantities. The EWGA, EWO and GFA of Apartment blocks show the best linear relationship with the total construction costs out of all building quantities. The BUA, EA and GBV of Apartment blocks show a more fluctuant relationship with the total construction costs. The GRA, IWGA and IWO of Apartment blocks show a poor linear relation and can be unreliable predictors. Withal, not all building quantities of other types of buildings have the same linear relationship with the total construction cost as can be seen in Appendix B. The data preparation (section 4.1.4) will take this into account.

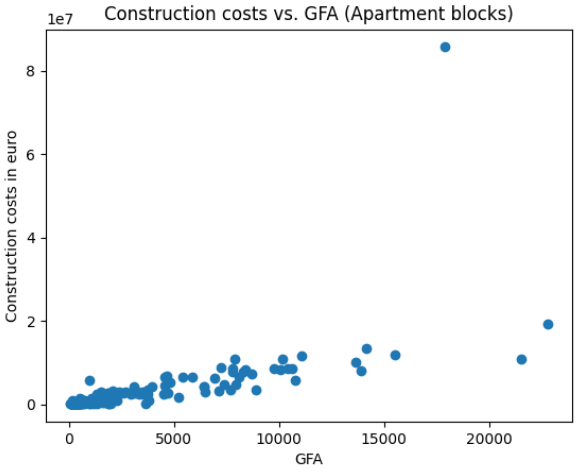

Construction costs vs. BUA (Apartment blocks)
1e7
Construction costs in euro
BUA

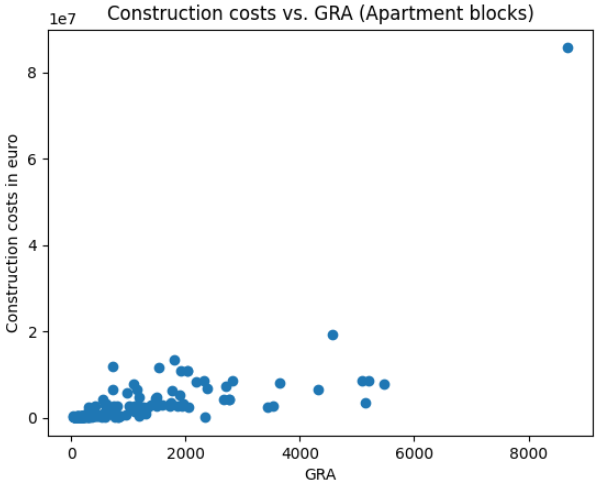

Construction costs vs. EA (Apartment blocks)
1e7
Construction costs in euro
EA

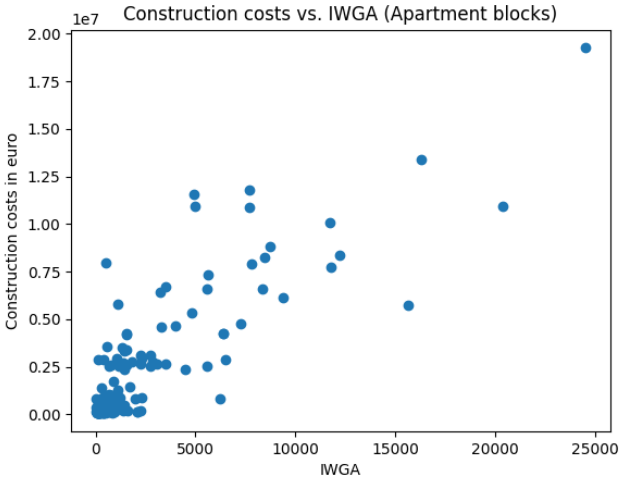

Construction costs vs. EWGA (Apartment blocks)
1e7
Construction costs in euro
EWGA

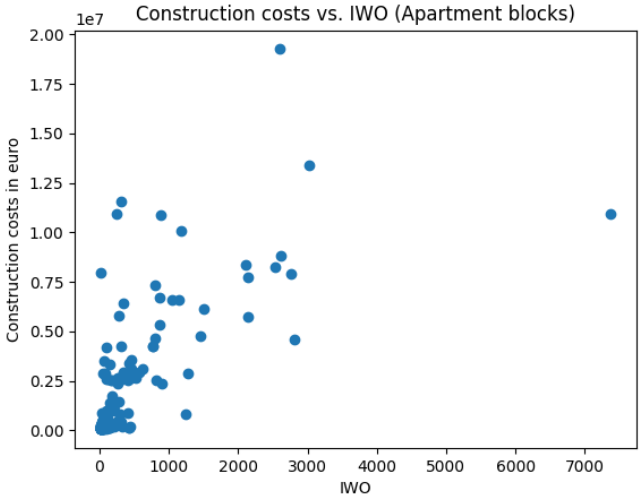

Construction costs vs. EWO (Apartment blocks)
1e7
Construction costs in euro
EWO


Construction costs vs. GBV (Apartment blocks)
1e7
Construction costs in euro
GBV


Construction costs vs. GFA (Apartment blocks)
1e7
Construction costs in euro
GFA







4.1.4 Data preparation

In section 4.1 the data used for the eventual model was analysed to identify characteristics and numbers in the data. 9 building quantities were analyzed, to recapitulate these building quantities are GFA, GBV, BUA, EWGA, EWO, GRA, EA, IWGA and IWO.

As remarked in section 4.1.3, the relationship between building quantities and the total construction costs for an individual type of project does not represent the relationship for all types of projects. However, pre-testing the selected machine learning model, which will be motivated in the next section, turned out that using all building quantities as input gave around the same result as selecting only a few building quantities with high availability and linear characteristics. This is probably due to the high correlation between building quantities or the absence of a linear relationship between certain building quantities and the construction costs, which causes the model to add a low weight to these input variables. Selecting all building quantities would result in more exclusion of data since not all building quantities for every project are available. Consequently, the most promising building quantities will be selected, based on linearity and included in the data to train and test the model. The selection of the different building quantities is shown in Table 11.

	GFA	GBV	BUA	EWGA	EWO	GRA	EA	IWGA	IWO	GIA
Apartment blocks	x			x	x					
Terraced houses	x	x		x						

Table 11: Selection of building quantities.

Furthermore, some data was removed from the dataset if the data complied with one of the following two criteria. First, the project was removed if the selected building quantities for a certain project were not all available or if the total construction costs were not available. The reason for removal was that if one variable was incomplete, often more variables were missing. The second criterium for exclusion was when a project was not realistic. More specifically, this means that the building showed characteristics that did not comply with a new construction. A common error was that renovations were included in the data, which led to many excluded projects. Table 12 shows the processing of the data and the exclusion of projects based on the two criteria mentioned in this section.

	Apartment blocks projects	**Terraced houses projects**
Initial number of projects	261	214
Excluded due to incompleteness	10	15
Excluded due to unrealistically	167	106
Total included	**84**	**93**

Table 12: Exclusion of data.

4.2 Model development

This section concerns the development of the model and elaborates on the methodology used for designing the machine learning model. Section 4.2.1 selects a model suitable for the data that was processed in the previous section. Section 4.2.2. elaborates on the methodology used in the machine learning model, which is demonstrated in section 4.2.3.

4.2.1 Model selection

There is not one single machine learning model suitable for every situation. Pre-processing turned out that there is a linear relation between the construction costs and some building quantities. Nevertheless, the pattern of the data is not the only feature that is important for selecting a model as becomes clear in section 3.2. Besides, it is difficult to compare all features of the different machine learning models. Therefore, this section focusses on trial-and-error. The models explained in section 3.2 are pre-tested with the Python model that is elaborated in the next sections. Table 13 yields the percentage of predictions that is within the 20% error range, which is the objective for a solution in this thesis.

	Apartment blocks	Terraced houses
Linear regression	38%	43%
Support vector machines	38%	46%
K-nearest neighbor	47%	40%
Decision tree	42%	39%
Random forest	48%	43%
Neural network	38%	40%

Table 13: Pre-testing machine learning types

Pre-testing shows that the percentage predictions differ between the several machine learning types. The machine learning model that produced the best predictions for apartment blocks and terraced houses is respectively random forest and support vector machines. The outcomes of these types will be explained more extensively in the next sections.

4.2.2 Design method

The goal of the design is to create a machine learning model that can perform cost estimations. There are several methods that can create a machine learning model. In this thesis, the machine learning model is created with Python. Python is a programming language, that is characterized as user-friendly, easily understandable and object-oriented. Python suits this thesis well, since the Python language is well interpretable. Another advantage of Python compared to other methods is that the python method is explainable. Some methods, like RapidMiner, only need an input and retrieve an output, based on your preferences. This decreases the ability to understand and explain how the model works. As Integrated Development Environment (IDE) PyCharm is used. An IDE allows to write, test and debug codes easier.

To design a machine learning model different packages are used. Packages are collections of Python modules that are pre-programmed and thus save time in programming. The packages that are used in this thesis are Pandas and Scikit-Learn. Pandas is used to create a fast and efficient data frame, besides, it has useful properties like a tool that can import data from excel. Excel is an often-used application and increases the accessibility of the machine learning design. Next to pandas, which is mainly used to import and export data, Scikit-Learn is used to actually process the data. Scikit-Learn is a simple and efficient tool for predictive data analysis. These machine learning models are pre-programmed and accessible to everyone. Scikit-Learn offers machine learning models varying from supervised learning to unsupervised learning and offers all machine learning models described in section 3.2. Therefore, Scikit-Learn is a useful package to use for designing this machine learning model.

4.2.3 Demonstration

This section describes the python code of the machine learning model in detail. The overall code is split and each part of the total code is explained. The overall code consists of importing the packages (section 4.2.1), importing the prepared data (section 4.1.4), splitting the data (section 3.1.3.2), training the model and validating the model. Different machine learning model types are created, however, all machine learning model types follow the same basic structure and only differ in the fact that another package is imported from Scikit-Learn that corresponds to the right machine learning model type. This section takes the multiple linear regression type to demonstrate the machine learning method.

```
#Import packages
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
```
Figure 8: Import packages.

Figure 8 shows that the packages are important. Pandas, which is used for the data frame, is shortened to "pd". Furthermore, from Scikit-Learn, that in Python is denoted as sklearn, the machine learning type package is imported. Also, from Scikit-Learn a package is important that is used for splitting the prepared data into a test and training set.

```
#Load the data
df = pd.read_excel("Test.xlsx")
predictors = df[['GFA','GBV','EWGA']]
predicted = df[["Costs"]]
```
Figure 9: Load the data.

After the packages are imported, the data is loaded for the machine learning model, as can be seen in Figure 9. A data frame is created, denoted in the code as "df". Pandas is used to create this data frame and import the data from excel. Furthermore, the data frame is split into predictor variables and the predicted variable. The predictor variables consist out of selected building quantities, whereas the predicted variable is logically the construction costs.

```python
#Create lossfunction Dataframes
aperesults_df=pd.DataFrame()

iterations = 2
for j in range(iterations):

    #Split
    X_train, X_test, y_train, y_test = train_test_split(predictors, predicted, test_size=0.20, shuffle=0)
    actual = y_test.to_numpy()

    #Train model
    reg = RANSACRegressor()
    reg.fit(X_train,y_train)

    #Predict
    prediction = reg.predict(X_test)

    #Validate
    Numbertest = range(35)
    for i in Numbertest:
        APE = abs((actual[i] - prediction[i]) / actual[i] * 100)
        PE = ((actual[i] - prediction[i]) / actual[i] * 100)

    #Load loss function into Pandas Dataframe
        df_resape=pd.DataFrame(data={"APE":APE})
        aperesults_df = aperesults_df.append(df_resape)
        lossfunctionape = aperesults_df[["APE"]].describe()
```

Figure 10: Train and validate model.

With the loaded data the model can be trained, as can be seen in Figure 10. First, the data is split into a training and test set. The Scikit-Learn package splits the predictor and predicted variables into four entities: the training predictor variables, the training predicted variable, the test predictor variables and the test predicted variable. The first two entities are used to train the model. Afterwards, a set of predictions is made using the test predictor variables entity. The predicted values are compared with the actual values that are retrieved from the test predicted variable entity. The results are stored into a data frame. This process is repeated multiple times, each time storing the results in the data frame.

4.3 Results

This section describes the results of the machine learning model of the previous section. The machine learning model has to be generalizable and therefore section 4.3.1 explains the methodology to validate the machine learning model. 4.3.2 uses these validation techniques and describes the outcome of the machine learning model, focusing on the error statistics.

4.3.1 Validation method

For the model validation, the MAPE is selected as loss function (section 3.1.3.1). The MAPE is a suitable validation method since it is relative to the actual construction costs. Loss functions, such as MAE, MSLE,

MSE are useful functions to consider when the range of the construction costs is small. However, if the range is wider these methods result in less interpretable outcomes since they are not compared to the original value. A difference of a million euros in a project of 200 million euros can be considered as a small difference, while the same difference in a project of one million euro's is high. Methods that are not compared with the actual costs are not capable of keeping this into account.

Furthermore, to check if the model is biased, cross-validation is applied to the model. The model is split into a training set and a test set. The type of cross-validation that was used is random subsampling. This type of cross-validation is not restricted to a limited number of iterations. At the same time, if the cross-validation results are more accurate, the model has a better fit. This better fit results in a less biased and more generalizable machine learning model. Previous studies have shown that it is not recommendable to have a test set outside the range of 10-30% (Eyo & Abbey, 2021). An often-used balance between training and test data is 80/20 (Bronshtein, 2021). Hence, this balance is used for the validation of the machine learning model. To prevent too much variation when validating 200 iterations are performed. This reduces the standard error since a higher number of validations is performed. Due to random subsampling, this high number can be achieved.

4.3.2 Outcome

4.3.2.1 Apartment blocks (internal entrance)

First, the machine learning model, with random forest as machine learning type, was trained with only data from the apartment blocks. Table 14 describes the percentage error of the apartment blocks projects between the predicted value and the actual value. In total data points were tested 3200 times in the machine learning model, which was trained 200 times with different randomized training data. In total 16 projects were tested in each iteration. The Mean Percentage Error (MPE) is -13.19% while the median (50%) is -1.59%. This means that there are stronger negative outliers than positive outliers. It can be concluded that the machine learning model tends to predict rather too high than too low.

Figure 11 confirms that negative outliers represent a bigger share of the outliers than positive outliers. Furthermore, the figure shows a normal distribution, which is centralized closely to the zero error. This points out that most predictions are closely predicted to the zero error.

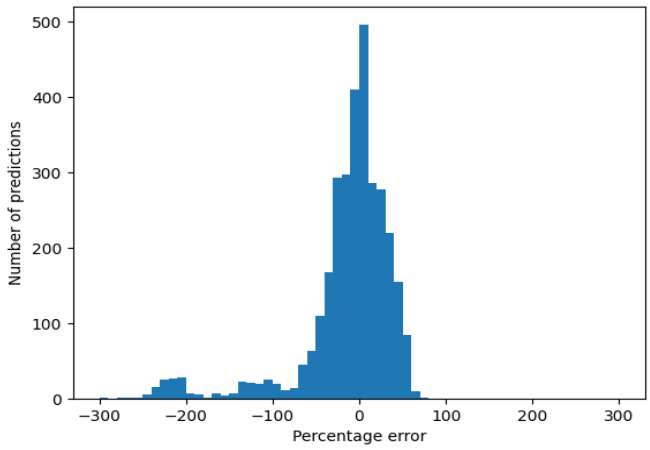


Figure 11: Percentage error apartment blocks.

Percentage error	
Count	3200
Mean	-13.19
Standard deviation	54.92
Min	-294.51
25%	-25.40
50%	-1.59
75%	17.51
Max	70.88

Table 14: Statistics percentage error apartment blocks

Percentage error describes the characteristics of the machine learning predictions well. However, for the interpretation of results, it is more convenient to analyze the absolute percentage error since the absolute percentage error does not distinguish positive or negative deviation. Table 15 shows the statistics of the absolute percentage error and Figure 12 visualizes the absolute percentage error. The best prediction deviated only 0.01% and as mentioned the worst prediction deviated 294.51%. Furthermore, 50% of all predictions deviated within the interval of 0 to 22.15%. The MAPE was 34.02%. Also, almost all predictions are within the 50% error range and there are barely very high outliers.

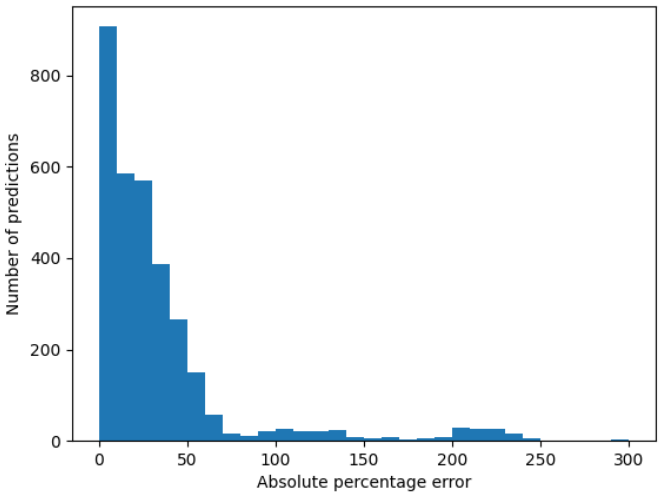


Figure 12: Absolute percentage error apartment blocks.

Absolute percentage error	
Count	3200
Mean	34.02
Standard deviation	45.08
Min	0.01
25%	8.58
50%	22.15
75%	38.89
Max	294.51

Table 15: Statistics absolute percentage error apartment blocks.

4.3.2.2 Terraced houses

Secondly, the machine learning model was trained with data from the terraced houses, this time using a support vector machine. Table 16 gives the statistics of the percentage error of terraced houses projects. The number of tested projects is just like the apartment blocks 3400 in 200 iterations. For the terraced houses, the machine learning model was again predicting too high, but the median of the percentage error was only -0.71%. The interquartile range, that is the range between the first quartile and third quartile, showed a comparable result compared to the apartment blocks. Nevertheless, the standard deviation is higher for the terraced houses in comparison with the apartment blocks. It can be concluded that the terraced houses have higher outliers than the apartment blocks. The maximum percentage error and minimum percentage error are relatively high compared to the apartment blocks, which enhances this conclusion. Also, the MPE is -13.26% and in combination with the median, it can be concluded that also this machine learning model suffers from negative outliers, rather than positive ones. Figure 13 strengthens this conclusion since higher negative percentage errors represent most of the total higher percentage errors. Also, the figure is characterized by a normal distribution, which is closer to the zero error.

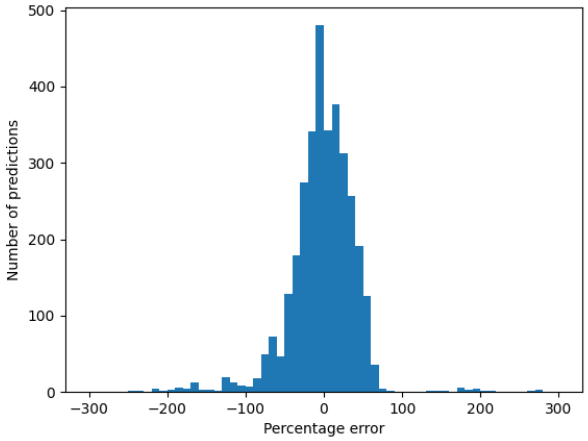


Figure 13: Percentage error terraced houses.

Percentage error	
Count	3400
Mean	-13.26
Standard deviation	111.99
Min	-1171.79
25%	-22.14
50%	-0.71
75%	24.08
Max	361.48

Table 16: Statistics percentage error terraced houses

Table 17 shows the statistics of the absolute percentage error of terraced houses. Outstanding is a MAPE with a value of 41.66% in combination with the third quartile (75%). The third quartile of the absolute error is still deviating heavily from the mean. This means that within the third quartile, the values are predicted with a lower percentage error, however, the error in the last quartile must be so high, that it causes a drastic increment of the MAPE. It can be concluded that the outliers in the dataset of the terraced houses are dominant to the model. Figure 14 shows the distribution of the absolute percentage error of the terraced houses. It also shows that disregarding the fact that the model suffers from some outliers, most of the predictions are within the 50% error range.

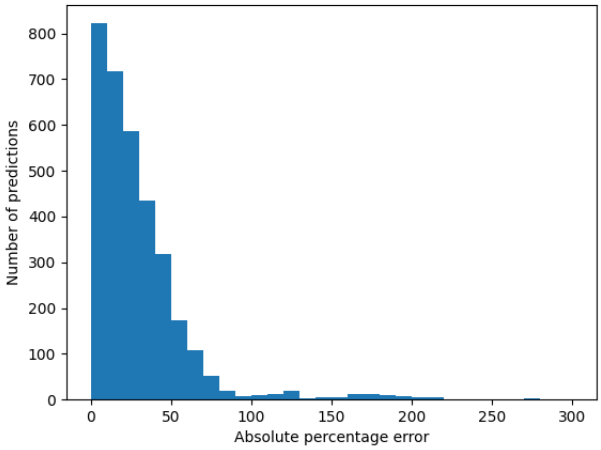


Figure 14: Absolute percentage error terraced houses.

Absolute percentage error	
Count	3400
Mean	41.66
Standard deviation	104.80
Min	0.01
25%	10.38
50%	22.86
75%	39.78
Max	1171.79

Table 17:Statistics absolute percentage error terraced houses.

4.4 Conclusion

In this conclusion, the following question is answered: *"How can machine learning facilitate a more accurate estimation of construction costs?"*

All building quantities showed a high availability and the construction costs have been defined more extensively. The building quantities have been pre-processed. Non-residential construction projects are divided into many classes, which resulted in fewer data per class. Therefore, to show the possibilities of machine learning, residential construction projects were used. Eventually, two classes with the most data were selected, namely terraced houses and apartment blocks. All nine building quantities were scatter plotted against the construction costs for these two classes. The only clear pattern that could be detected was a linear relationship for some building quantities. For both classes, the building quantities with the most linear relationship were selected. Afterwards, the data was cleaned and if the corresponding selected building quantities of a class were not present this project was excluded from the data. Furthermore, if a project showed unrealistic features, that could not represent a new construction, the project was excluded as well from the data.

Several characteristics of a machine learning type determine if the type is suitable for the prepared data. Pre-testing selected the best machine learning types for apartment blocks and terraced houses based on the number of predictions that were within a 20% error range. Python is used as the programming language since it is user-friendly, easily understandable and object-oriented. PyCharm is used as IDE to write, test and debug codes easier. Besides, several packages are imported to support the programming process.

The MAPE is selected as loss function to validate the results since it is relative to the actual construction costs. The type of cross-validation that was used is random subsampling because it is not restricted to a limited number of iterations. The balance between test and training data was 20/80 and 200 iterations were performed. The machine learning predictions of apartment blocks and terraced houses resulted in a MAPE of 34.02% and 41.66% respectively. However, statistics showed the model suffered from outliers. The median absolute percentage error of the machine learning predictions of apartment blocks and terraced houses resulted in 22.15% and 22.86% respectively.

5 Evaluation

The last section contained the design & development and demonstration stage of the DSRM process. The fifth step in DSRM is to observe and measure how well the model offers a solution for the problem. Section 2 offered an objective for a solution, which will be used in this section to evaluate the model.

The previous section showed the objective statistics of the results of the machine learning model. This section heads more to a subjective interpretation of the results. Section 5.1 shows the interpretation of the results, whereas Section 5.2 contributes to the practical implications. Section 5.3 shows the limitations of the research.

5.1 Evaluation of results

The aim of this thesis was to improve the conventional cost estimation methods in the program phase of the construction process. The literature revealed a good perspective for machine learning in the construction sector and even proved that some machine learning predictions could get close to the actual value. The machine learning predictions of apartment blocks and terraced houses yielded a MAPE of 34.02% and 41.66% respectively. The error was not expected and a far from the optimal result. However, the median of the absolute percentage error of the predictions yielded 22.15% and 22.86% respectively. This was already a more positive result since it shows that the MAPE is mostly influenced by outliers. Analyzing these outliers is important to be able to improve in further research. The outliers that caused the highest errors were mostly data points with low construction costs. However, the great majority of the data was predicted within the percentage error range of 50%.

The deviation from the promising results in previous studies can have several causes. The deviation can be caused by the amount of construction costs components. It can be the case that some cost components are more influenced by certain building quantities or that some costs cannot be retrieved from building quantities. One respondent strengthens this belief and pointed out that for example, the difference in cost between central heating and a heat pump is high. This type of installation cost is a typical example of a cost component that is difficult to predict by building quantities. Input variables that could predict specific cost components would be useful in this case.

The machine learning model yields a solution for the problem if most of the predictions are within the 20% error range, as was stated in the objective for a solution in Section 2. Respectively 46% and 48% of all predictions of the terraced houses and apartment blocks were within a 20% error range. Taking into account that almost half of the predictions are within the 20% error range, this thesis offers promising results for further research. Literature supported the choices made in this thesis and offered promising results for machine learning predictions in the construction sector. Together, with other research limitations research, this thesis offers an advantage for further research in predicting construction costs with machine learning.

5.2 Evaluation practical implications

Currently, cost estimation is mostly performed in an analogous way, relying on earlier comparable projects. Machine learning is a technique that is not widely adopted in the construction sector. This technology focuses more on the relationship between variables and could cause a shift from analogous estimation towards parametric estimation. However, first, the benefits of the machine learning approach must be proven. Machine learning offers the advantage of lower expertise needed to perform cost estimations when only some building quantities are known. This could decrease the expertise threshold that is currently needed to perform cost estimations. Besides, key cost figures that are used at the moment can become unnecessary, because calculations do not have to be performed anymore. Nevertheless, accuracy is by far

the most important factor considered when making the choice for machine learning. Even though earlier studies show already good potential, the analogous approach has definitely its advantages compared to machine learning. One advantage is the expertise and experience in cost estimation. Cost experts can more easily point out expensive or uncommon parts in the construction process, whereas a machine learning model needs a specific input that describes these expensive or uncommon parts. The ability to deal with uncommonness is a limitation of the machine learning model since it only has three input variables and cannot deal well with variation.

5.3 Research limitations

The desired accuracy of the machine learning model was not met in this thesis. Therefore, it is important to mention the limitations of this research. This way, this thesis can prevent limitations in further research and it can be used as a fundament for new research. The following limitations were encountered:

- The data used for the machine learning model was low quality. This thesis focused on predicting the construction costs for residential constructions in the program phase of the construction process. The data was all remarked as new constructions, however, reviewing these projects with a cost expert turned out that there were often wrong annotations in the dataset. This led to the exclusion of a great share of the projects, that could not be used for making predictions. More projects lead in general to better predictions.
- The selected methodology in this thesis offered some limitations as machine learning is a technique that requires trial-and-error. The methodology in this thesis gave the opportunity to evaluate and improve after each cycle, but machine learning requires quick trial-and-error rather than cycles. The methodology for designing the machine learning model is theoretical, but a more practical methodology would probably fit better. The preferred situation would be to find a suitable machine learning model and afterwards explore why it works, instead of a theoretical framework that makes the assumption that it works on your model. This thesis used pre-testing to tackle this problem, but is in general a limitation.
- The unstructured interviews were all conducted at the same company. The employees of this company worked probably all according to the same protocol since the answers of the respondents did not fluctuate much. The fact that the company is a corporate organisation makes it even more likely that there is a standardized way of working. The risk of performing unstructured interviews this way is that it is not generalizable for the whole construction sector and therefore yields a bias.
- The data used to train and test the machine learning model was solely originating from Dutch construction projects. Also, in this thesis, the construction phases are defined in a Dutch context, which makes it difficult to generalize since the construction sector is a worldwide sector. Therefore, this thesis should be seen in the context of the Dutch construction sector.

6 Communication

This section communicates the results of this thesis. Section 6.1 is dedicated to answering the research question by using the sub-questions from different sections. Since this thesis was executed at Arcadis, practical recommendations that follow from this thesis are given in section 6.2. Section 6.3 finishes this thesis with further research recommendations.

6.1 Conclusion

This section answers the questions formulated at the beginning of this thesis. Throughout this thesis, conclusions have been drawn in sections that contained a sub-question. Therefore, the sub-questions will be recapitulated and answered shortly. The conclusions of these sub-questions will support the conclusion of the main question. Section 6.1.1 answers the sub-questions and section 6.1.2. will finish giving an overall conclusion of this thesis, answering the main question.

6.1.1 Conclusion sub-questions

To retrieve a norm for the current possibilities in cost estimation the first section analyzed the research context. The following questions are answered:

1. *What are the current cost estimators in the program phase in literature?*
    a. *What cost-related data is available during the program phase?*
    b. *How accurately are conventional prediction methods in the program phase?*

SQ1a: The program phase consists out of three subphases. In chronological order, these are the concept phase, feasibility phase and project definition. The concept phase describes the reason for a new building and no cost-related data is available. On the other hand, in the feasibility phase and project definition fundamental ideas are denoted and a statement of requirements is set up. Especially the object-related part of the statement of requirements is useful since it includes a wide range of information and building quantities can be derived from this document.

SQ1b: The accuracy of cost estimation in the program phase depends on the complexity and the nature of the building. A unique and special building can be estimated less accurate and a cost estimation takes more time. More standard and simple buildings, such as houses and offices, can be estimated with a percentage error between 10-20%, whereas more complex buildings can be estimated with an error between 20-30%. The most time-intensive process is in extracting information out of the statement of requirements.

SQ1: Based on literature and the conducted interviews it is most convenient to conclude that the conventional cost estimation method is a combination of analogous estimation and expert judgement. The information for new construction is extracted from early constructions. The major components of previous similar jobs are taken into account when estimating the costs for a new building. A drawback of this method is that know-how is required and enough comparable projects are needed.

In order to construct a reasonable machine learning model that is capable of predicting construction costs, the literature was consulted. The next question is answered:

2. *What types of machine learning can be applied to estimate construction costs?*

SQ2: Machine learning is in literature acknowledged as the study of computer algorithms that seek to improve automatically through experience. There are several approaches in machine learning, that can be divided into 4 types of machine learning. The approach depends on the type of value that is predicted, which can be either a continuous or discrete value. Besides the type of value, the approach depends on if the data is labelled or not. Based on the aim of this thesis, the machine learning approach in this thesis can be defined as a regression problem. There is not one machine learning type suitable for all types of predictions. Factors like the implementation speed, ease of interpreting results, nature of supervised machine learning task, nature and amount of input and output data and the complexity of the model to be learnt are crucial factors to consider when deciding on a machine learning algorithm. It can be concluded that the choice for a machine learning type depends on multiple factors.

As can be concluded from the literature section, there is no one factor that is decisive for the selection of a machine learning type. An important factor is the features of the cost-related data. Also, the design requires a proper method to process the data. Consequently, these questions are answered:

3. *How can machine learning be applied to cost-related data in the program phase for cost estimation?*
    a. *What are the features of the cost-related data?*
    b. *How can a machine learning process to the data?*

SQ3a: The cost-related data that is suitable for prediction are building quantities derived from the statement of requirements. Both the input and the output are continuous values. Some building quantities show a better degree of linearity with the construction costs compared to other building quantities. Also, the construction projects showed a wide range of construction costs. Since the dataset showed inappropriate data points, a data preparation was executed before implementing it in the machine learning model.

SQ3b: There is not one machine learning type suitable for each situation. Each machine learning type has its own advantages and drawbacks. Therefore, the discussed machine learning types were pre-tested to select a machine learning type. For the apartment blocks, this resulted in the random forest machine learning type and for the terraced houses, the support vector machine was superior.

SQ3: The model was created using Python as a programming language. Using this programming language, the packages of Sci-kit Learn were applied to the prepared data from the program phase of the construction to train the model. The test prepared test set was used to validate the model. The model was validated using cross-validation and the error was measured with the MAPE loss function.

After the model was analyzed and the model was developed. The data was applied to the model and gave results that could be compared with the norm. The results needed objective interpretation using summary statistics. The following question was answered:

4. *What are the results from cost estimation using the created machine learning model?*

SQ4: The machine learning predictions of the apartment blocks and terraced houses showed quite similar results. The MAPE of the predictions of the apartment blocks and terraced houses were 34.02% and 41.66%

respectively. The median absolute percentage error of the machine learning predictions of apartment blocks and terraced houses resulted in 22.15% and 22.86% respectively. The model suffers slightly from outliers since the mean is deviating from the median, however, this difference is mostly caused by a couple of high outliers since almost all predictions are within the 50% error range.

6.1.2 Conclusion main question

The previous section answered the sub-questions. This section gives the overall conclusion based on the conclusions of the different sections throughout this thesis and the answers to the sub-questions. The main question of this thesis was formulated as follows:

How can machine learning improve cost estimation in the program phase of the construction process?

Main question: The objective of this thesis was to surpass the conventional cost estimation methods in the program phase of the construction process. The current method was using a combination of analogous estimation combined with expert judgment. The proposed method in this thesis, machine learning, can be classified more as parametric cost estimation. The conventional cost estimation method had an error of 10-20% for rather simplistic buildings, such as offices and houses. The data of two types of construction projects were used to perform predictions, namely terraced houses and apartment blocks. Due to pre-testing, two different machine learning types were selected for the two building classes.

The results of the machine learning model were inaccurate. Whereas the percentage error of the current estimation method for the predicted classes lays between 10-20%, the machine learning model had for most predictions a higher percentage error. The absolute percentage error distribution shows that 46% of the total predictions of the terraced houses are within the 20% error range. For the apartment blocks, 48% of the total predictions were within the 20% error range. Since in most cases the machine learning model does not surpass the accuracy of the conventional method, it can be concluded that this thesis does not show that machine learning cost estimation method can improve the conventional cost estimation method.

6.2 Practical recommendations

This thesis does not prove that the accuracy is higher compared to conventional methods. The machine learning model that was created in this thesis has to be improved if it potentially could work out as a cost estimator in the program phase of the construction process. At the same time, machine learning is growing and the technology is more adapted in the world. Whereas machine learning is currently a technology that is not widely adopted by the construction sector, it might be in 10 or 20 years. It is important to keep an eye on the developments and if there is a generally adopted machine learning strategy that can outperform conventional methods, this can yield a competitive advantage.

Besides, data is getting more important and available in the 21st century. A lot of actions we perform every day are data-driven and it is not expected that the importance of data will decrease soon. In addition, the generation and use of data will probably only grow in the coming years. The construction sector generates a vast amount of data that is often not even processed. As seen in this thesis, machine learning has the ability to handle a lot of data and it will even get stronger when it uses more data. The prospect for the future and the way machine learning works results in the undeniable fact that capturing data is getting more important. The right administration of data gives the machine learning model the opportunity to keep improving.

6.3 Research recommendations

Recommendations for further research in the field of construction cost prediction with machine learning:

- Adapt a methodology that allows making quick improvements on the machine learning model. Machine learning is a technique that requires a lot of trial-and-error. This thesis shows that a good theoretical argumentation gives no guarantees for a good performing machine learning model. Big improvement cycles are too time expensive and preferably these should be short improvement cycles.
- Integrate international construction projects in further research. This could result in a model that is internationally applicable. These projects might not have the same characteristics as Dutch construction projects, but it is worth knowing the characteristics of these projects to create a wider machine learning acceptance in the construction sector.
- Predict the construction costs more specific. This thesis tried to predict the overall construction costs of a building, however, the machine learning model could hardly make distinctions between certain cost components. Giving the machine learning model sub-costs as input could potentially result in a more accurate prediction, since the model can recognize the relation between the building quantities and the individual cost components.

7 References

Agarwal, R., Chandrasekaran, S., & Sridhar, M. (2016). *Imagining construction's digital future*.

Ali, H., Eldrup, N. H., Normann, F., Skagestad, R., & Øi, L. E. (2019). *Cost Estimation of CO 2 Absorption Plants for CO 2 Mitigation-Method and Assumptions*. https://doi.org/10.1016/j.ijggc.2019.05.028

Alshamrani, O. S. (2017). Construction cost prediction model for conventional and sustainable college buildings in North America. *Journal of Taibah University for Science*, *11*, 315–323. https://doi.org/10.1016/j.jtusci.2016.01.004

Arcadis. (2021). *Being an Arcadian*. https://www.arcadis.com/en/become-an-arcadian/being-an-arcadian

Azlinah, M. W. B., Bee, M., & Yap, W. (2020). *Supervised and Unsupervised Learning for Data Science Unsupervised and Semi-Supervised Learning Series Editor: M. Emre Celebi*. http://www.springer.com/series/15892

Bao, W., Lianju, N., & Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems With Applications*, *128*, 301–315. https://doi.org/10.1016/j.eswa.2019.02.033

Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., Owolabi, H. A., Alaka, H. A., & Pasha, M. (2016). *Big Data in the construction industry: A review of present status, opportunities, and future trends*. https://doi.org/10.1016/j.aei.2016.07.001

Bronshtein, A. (2021). *Train/Test Split and Cross Validation in Python*. https://resources.experfy.com/bigdata-cloud/train-test-split-and-cross-validation-in-python/

Brownlee, J. (2019). *How to Choose Loss Functions When Training Deep Learning Neural Networks*. https://machinelearningmastery.com/how-to-choose-loss-functions-when-training-deep-learning-neural-networks/

Carl Carande, Paul Lipinski, & Traci Gusher. (2017). *How to Integrate Data and Analytics into Every Part of Your Organization*. https://hbr.org/2017/06/how-to-integrate-data-and-analytics-into-every-part-of-your-organization

Chakraborty, D., Elhegazy, H., Elzarka, H., & Gutierrez, L. (2020). *A novel construction cost prediction model using hybrid natural and light gradient boosting*. https://doi.org/10.1016/j.aei.2020.101201

Choi, J., Gu, B., Chin, S., & Lee, J.-S. (2019). *Machine learning predictive model based on national data for fatal accidents of construction workers*. https://doi.org/10.1016/j.autcon.2019.102974

Cisco. (2018). Annual internet report (2018-2023). In *White paper Cisco public*. https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.pdf

Cooper, D. R., & Schindler, P. S. (2014). Business Research Methods 12th Edition. In *Business Research Methods*.

Czajkowski, M., & Kretowski, M. (2019). Decision tree underfitting in mining of gene expression data. An evolutionary multi-test tree approach. *Expert Systems With Applications*, *137*, 392–404. https://doi.org/10.1016/j.eswa.2019.07.019

Dashore, A. (2021). *Methods of Cost Estimation in Projects - Tools and Techniques - Construction Engineering & Management*. https://theconstructor.org/construction/methods-cost-estimation/36532/

Desai, M., & Shah, M. (2020). *An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and Convolutional neural network (CNN)*. https://doi.org/10.1016/j.ceh.2020.11.002

Dorafshan, S., & Azari, H. (2020). *Evaluation of bridge decks with overlays using impact echo, a deep learning approach*. https://doi.org/10.1016/j.autcon.2020.103133

Dubitzky, W., Granzow, M., & Berrar, D. (2007). *Fundamentals of Data Mining in Genomics and Proteomics*.

Elfahham, Y. (2019). *Estimation and prediction of construction cost index using neural networks, time series, and regression*. https://doi.org/10.1016/j.aej.2019.05.002

Ellingson, S. R., Davis, B., & Allen, J. (2020). *Machine learning and ligand binding predictions: A review of data, methods, and obstacles*. https://doi.org/10.1016/j.bbagen.2020.129545

Eyo, E. U., & Abbey, S. J. (2021). *Machine learning regression and classification algorithms utilised for strength prediction of OPC/by-product materials improved soils*. https://doi.org/10.1016/j.conbuildmat.2021.122817

Fenza, G., Gallo, M., Loia, V., Orciuoli, F., & Herrera-Viedma, E. (2021). Data set quality in Machine Learning: Consistency measure based on Group Decision Making. *Applied Soft Computing Journal*, *106*. https://doi.org/10.1016/j.asoc.2021.107366

Fildes, R., & Goodwin, P. (2021). Stability in the inefficient use of forecasting systems: A case study in a supply chain company. *International Journal of Forecasting*, *37*, 1031–1046. https://doi.org/10.1016/j.ijforecast.2020.11.004

Gerhard Unger, & Marc Rodt. (2019). *The Power of Algorithmic Forecasting*. https://www.bcg.com/publications/2019/power-of-algorithmic-forecasting

Ghatak, A. (2017). *Machine Learning with R*. https://doi.org/10.1007/978-981-10-6808-9

Gujarati, D. N. (2019). *Linear Regression: A Mathematical Introduction*. https://doi.org/10.4135/9781071802571

Hu, R., Chen, K., Chen, W., Wang, Q., & Luo, H. (2021). *Estimation of construction waste generation based on an improved on-site measurement and SVM-based prediction model: A case of commercial buildings in China*. https://doi.org/10.1016/j.wasman.2021.04.012

Huang, C.-H., & Hsieh, S.-H. (2020). *Predicting BIM labor cost with random forest and simple linear regression*. https://doi.org/10.1016/j.autcon.2020.103280

Hughes, R. T. (1996). judgement as an estimating method. In *Information and Software Technology* (Vol. 38).

Javatpoint. (2021). *Overfitting And Underfitting in Machine Learning*. https://www.javatpoint.com/overfitting-and-underfitting-in-machine-learning

Jawad, J., Hawari, A. H., & Javaid Zaidi, S. (2021). Artificial neural network modeling of wastewater treatment and desalination using membrane processes: A review. *Chemical Engineering Journal*, *419*, 129540. https://doi.org/10.1016/j.cej.2021.129540

Jeong, J. H., Woo, J. H., & Park, J. (2020). *Machine Learning Methodology for Management of Shipbuilding Master Data*. https://doi.org/10.1016/j.ijnaoe.2020.03.005

Kaviani, S., & Sohn, I. (2021). Application of complex systems topologies in artificial neural networks optimization: An overview. *Expert Systems With Applications*, *180*, 115073. https://doi.org/10.1016/j.eswa.2021.115073

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*, 1235–1242. https://doi.org/10.1016/j.buildenv.2004.02.013

Kubina, M., Varmus, M., & Kubinova, I. (2015). Use of big data for competitive advantage of company. *Procedia Economics and Finance*, *26*, 561–565. https://doi.org/10.1016/S2212-5671(15)00955-7

Lawson, C. E., Martí, J. M., Radivojevic, T., Vamshi, S., Jonnalagadda, R., Gentz, R., Hillson, N. J., Peisert, S., Kim, J., Simmons, B. A., Petzold, C. J., Singer, S. W., Mukhopadhyay, A., Tanjore, D., Dunn, J. G., & Martin, H. G. (2021). Machine learning for metabolic engineering: A review. *Metabolic Engineering*, *63*, 1096–7176. https://doi.org/10.1016/j.ymben.2020.10.005

Lester, A. (2017). *Project Management, Planning and Control*. https://doi.org/10.1016/B978-0-08-102020-3.00013-9

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, *16*(6), 321–332. https://doi.org/10.1038/nrg3920

Liu, H., Fu, Z., Yang, K., Xu, X., & Bauchy, M. (2019). *Machine learning for glass science and engineering: A review*. https://doi.org/10.1016/j.nocx.2019.100036

Lu, H.-J., Zou, N., Jacobs, R., Afflerbach, B., Lu, X.-G., & Morgan, D. (2019). *Error assessment and optimal cross-validation approaches in machine learning applied to impurity diffusion*. https://doi.org/10.1016/j.commatsci.2019.06.010

Maheshwera, U., Paturi, R., & Cheruku, S. (2020). *Application and performance of machine learning techniques in manufacturing sector from the past two decades: A review*. https://doi.org/10.1016/j.matpr.2020.07.209

Maleki, F., Ovens, K., Najafian, K., Forghani, B., Reinhold, C., & Forghani, R. (2020). *Overview of Machine Lear ning Part 1 Fundamentals and Classic Approaches*. https://doi.org/10.1016/j.nic.2020.08.007

Mangalathu, S., Shin, H., Choi, E., & Jeon, J.-S. (2021). Explainable machine learning models for punching shear strength estimation of flat slabs without transverse reinforcement. *Journal of Building Engineering*, *39*. https://doi.org/10.1016/j.jobe.2021.102300

Matel, E. (2019). *An artificial neural network approach for cost estimation of engineering services*.

Mazzi, F. (2021). CMOs and AI: Can trained machine learning be justified with the concept of know-how? *World Patent Information*, *65*. https://doi.org/10.1016/j.wpi.2021.102036

Nanda, G., Vallmuur, K., & Lehto, M. (2018). *Improving autocoding performance of rare categories in injury classification: Is more training data or filtering the solution?* . https://doi.org/10.1016/j.aap.2017.10.020

Osarogiagbon, A. U., Khan, F., Venkatesan, R., & Gillard, P. (2021). Review and analysis of supervised machine learning algorithms for hazardous events in drilling operations. In *Process Safety and Environmental Protection* (Vol. 147, pp. 367–384). https://doi.org/10.1016/j.psep.2020.09.038
Pasini, M. (2021). *A scalable algorithm for the optimization of neural network architectures*. https://doi.org/10.1016/j.parco.2021.102788
Peffers, K., Tuunanen, T., Rothenberger, M., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Lakartidningen*, *73*(48), 4201–4204.
Pijpers, I. R., & van der Woude, D. H. J. (2012). *Jellema 1: Bouwnijverheid*. https://www.scribd.com/doc/28228230/3/Bouworganisatievormen-naast-elkaar
Project Management Institute. (2016). Project Cost Management. In *Case Studies in Project, Program, and Organizational Project Management*. https://doi.org/10.1002/9780470549179.ch7
Scikit Learn. (2021). *Linear Models - Robustness regression: outliers and modeling errors*. https://scikit-learn.org/stable/modules/linear_model.html#robustness-regression-outliers-and-modeling-errors
Shakil, S., Arora, D., & Zaidi, T. (2021). *Feature based classification of voice based biometric data through Machine learning algorithm*. https://doi.org/10.1016/j.matpr.2021.05.261
Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2016). *Critical analysis of Big Data challenges and analytical methods-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)*. https://doi.org/10.1016/j.jbusres.2016.08.001
Stetco, A., Dinmohammadi, F., Zhao, X., Robu, V., Flynn, D., Barnes, M., Keane, J., & Nenadic, G. (2018). *Machine learning methods for wind turbine condition monitoring: A review*. https://doi.org/10.1016/j.renene.2018.10.047
Suresh, S., Sundararajan, N., & Savitha, R. (2013). *Supervised Learning with Complex-valued Neural Networks*. http://www.springer.com/series/7092
Swei, O., Gregory, J., & Kirchain, R. (2017). Construction cost estimation: A parametric approach for better estimates of expected cost and variation. *Transportation Research Part B*, *101*, 295–305. https://doi.org/10.1016/j.trb.2017.04.013
Tan, C. H., Siah Yap, K., & Yap, H. J. (2012). Application of genetic algorithm for fuzzy rules optimization on semi expert judgment automation using Pittsburg approach. *Applied Soft Computing*, *12*, 2168–2177. https://doi.org/10.1016/j.asoc.2012.03.018
Urbanowicz, R. J., & Browne, W. N. (2017). *Introduction to Learning Classifier Systems*. http://www.springer.com/series/11845
U.S. Bureau of Labor Statistics. (2021). *Cost Estimators : Occupational Outlook Handbook: : U.S. Bureau of Labor Statistics*. https://www.bls.gov/ooh/business-and-financial/cost-estimators.htm#tab-2
van Klompenburg, T., Kassahun, A., & Catal, C. (2020). *Crop yield prediction using machine learning: A systematic literature review*. https://doi.org/10.1016/j.compag.2020.105709
Vandeput, N. (2019). *Forecast KPIs: RMSE, MAE, MAPE & Bias | by Nicolas Vandeput | Towards Data Science*. https://towardsdatascience.com/forecast-kpi-rmse-mae-mape-bias-cdc5703d242d
Xu, Y., Zhou, Y., Sekula, P., & Ding, L. (2021). *Machine learning in construction: From shallow to deep learning*. https://doi.org/10.1016/j.dibe.2021.100045
Yigit, S. (2021). *A machine-learning-based method for thermal design optimization of residential buildings in highly urbanized areas of Turkey*. https://doi.org/10.1016/j.jobe.2021.102225

8 Appendix

8.1 Appendix A: Translation construction phases

Dutch main phase	*English translation*	*Dutch sub phase*	*English translation*
Programma	Program	**Initiatief**	Concept phase
		Haalbaarheidsstudie	Feasibility phase
		Projectdefinitie	Project definition
Ontwerp	Design phase	**Structuurontwerp**	Sketch design
		Voorlopig ontwerp	Preliminary design
		Definitief ontwerp	Definitive design
Uitwerking	Pre-building phase	**Bestek**	Technical design
		Prijsvorming	Pricing phase
Realisering	Realization	**Werkvoorbereiding**	Preparation phase
		Uitvoering	Execution phase
		Oplevering	Delivery phase
Beheer	Maintenance	**Beheer**	Maintenance
Sloop	Demolition	**Sloop**	Demolition

8.2 Appendix B: Unstructured interviews (summarized)

What information/data can you use to perform cost estimations in the program phase?
Respondent 1:
- The functional units are used to derive building quantities.
- The new type of building is referred to earlier projects.
- Certain types of buildings will stay within a bandwidth.

Respondent 2:
- Earlier comparable projects are consulted.
- Key cost figures are used for cost estimation.

Respondent 3:
- Statement of requirements gives the specifications of a building.

Respondent 4:
- Building quantities is the most important information
- Key cost figures are used to translate these to a cost estimation

How accurate can you estimate the construction cost using this information/data?
Respondent 1:
- Cost estimations can be done with a +/- 10% error.

Respondent 2:
- Depends on the accuracy of the information.
- Cost estimations can be done with +/- 20% error.

Respondent 3:
- Depends on the nature of the project.
- More standard buildings can be estimated more accurate.
- Cost estimations can be done with 10% error for houses and offices.

Respondent 4:
- Depends on the nature of the project.
- For rare and more complex buildings cost estimations have between 20-30% error.
- For standard buildings this is a 15% error.

How much time does the cost estimation in the program (Dutch translation: programma) take you?
Respondent 1:
- Depends on the references available.
- On average within one day.

Respondent 2:
- Depends on the type of project.
- Fluctuates depending on the project between a half day and two days.
- Reading the statement of requirements is a time-consuming process.

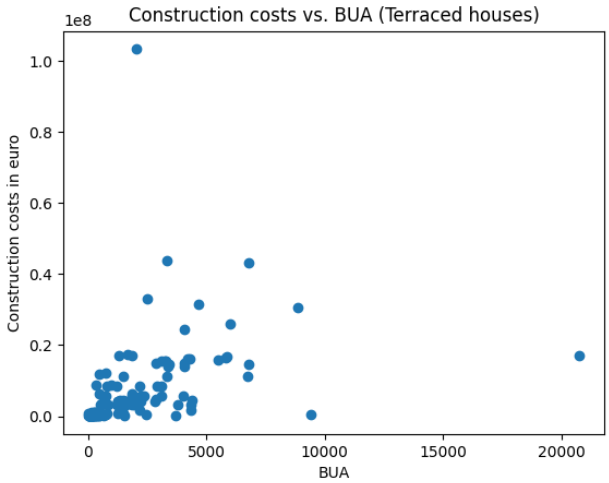
Respondent 3:

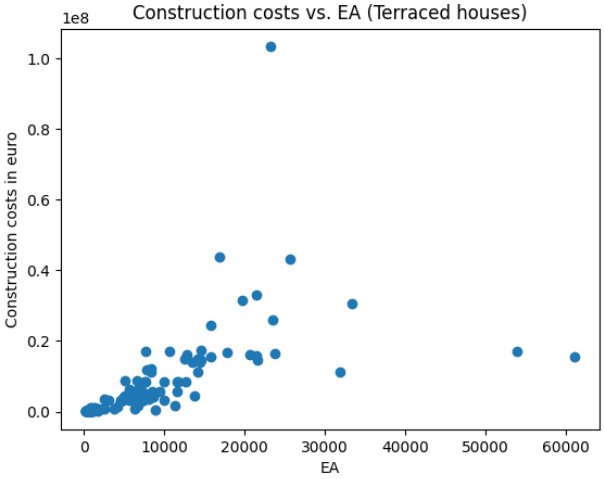
- Depends on the type of project.
- One hour for the respondent's specialization, but three days for a unique building.

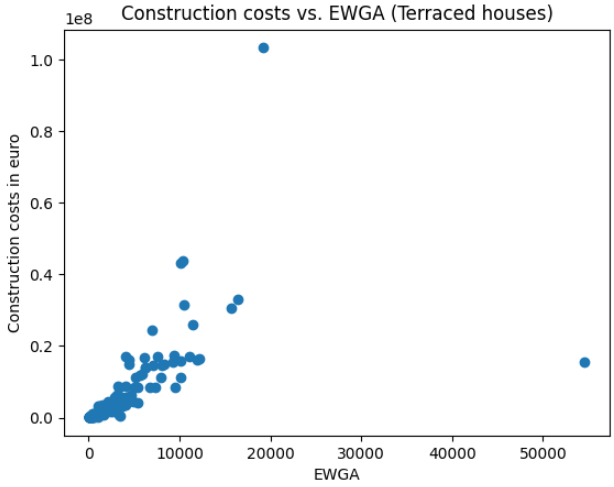
Respondent 4:

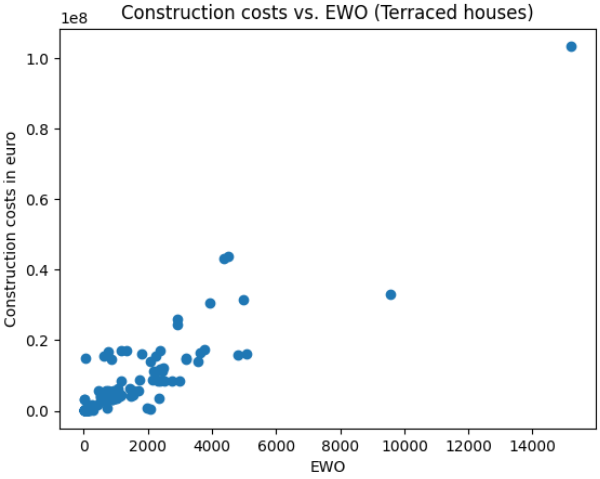
- Big, unique buildings are more time-consuming than standard constructions.

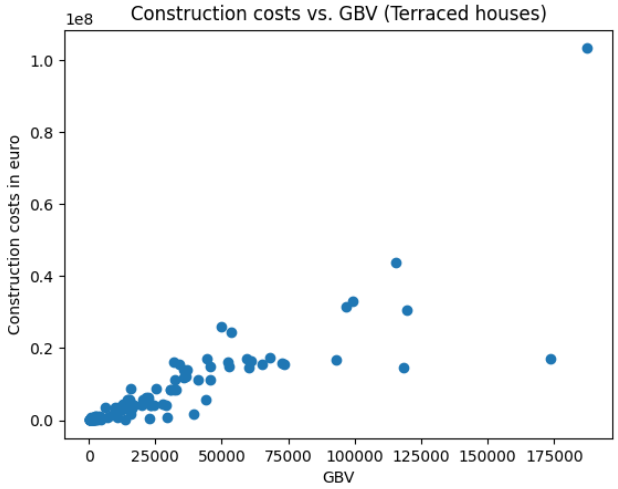
8.3 Appendix C: Scatter plot Terraced houses

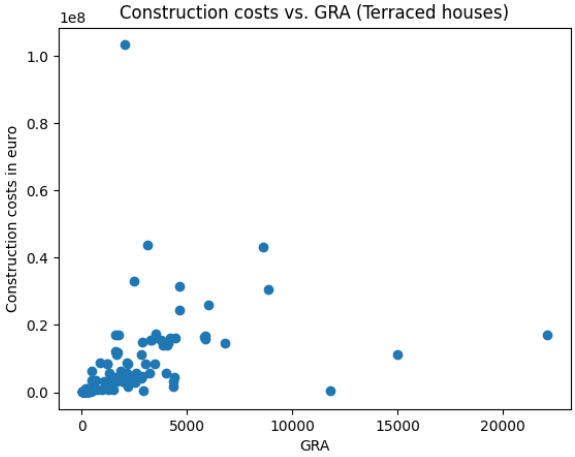


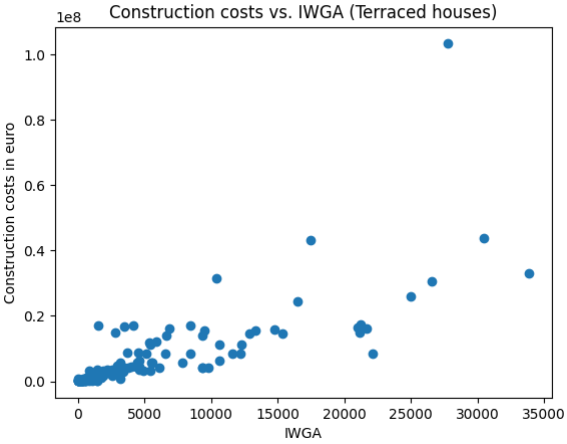


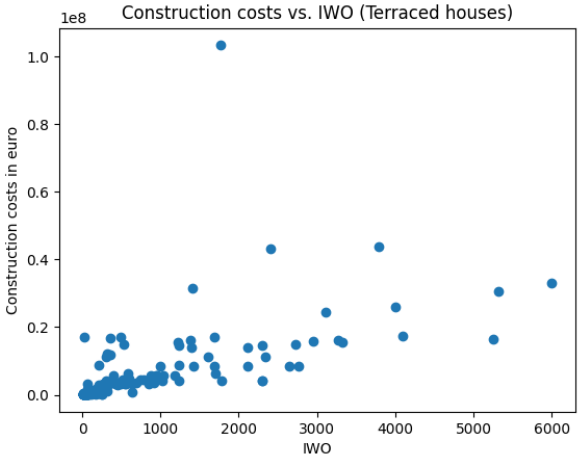







Construction costs vs. GRA (Terraced houses)
1e8
1.0
0.8
0.6
0.4
0.2
0.0
Construction costs in euro
0
5000
10000
15000
20000
GRA


Construction costs vs. IWGA (Terraced houses)
1e8
1.0
0.8
0.6
0.4
0.2
0.0
Construction costs in euro
0
5000
10000
15000
20000
25000
30000
35000
IWGA


Construction costs vs. IWO (Terraced houses)
1e8
1.0
0.8
0.6
0.4
0.2
0.0
Construction costs in euro
0
1000
2000
3000
4000
5000
6000
IWO