Department of Building Environment

Urban Systems & Real Estate

Improving the accuracy of project duration with artificial intelligence

- A case study within the utility construction industry

by:

T. (Tarik) Azaaj

Student number: 0974461

Supervisors:

prof.dr.ir. B. (Bauke) de Vries	*TU/e Urban Systems & Real Estate*
ir.ing. A. (Aant) van der Zee	*TU/e Urban Systems & Real Estate*
dr. T. (Tao) Feng	*TU/e Urban Systems & Real Estate*

in partial fulfilment of the requirement for the degree of:

Master of Science in Construction, Management and Engineering

Eindhoven, Monday 31th December, 2018

Abstract

This research focusses on the prediction of utility construction projects. These predictions are important for contractors because of growing costs. To improve the project duration of the utility construction projects, this research tests two different techniques through a case study. The case study is conducted at the utility department of a contractor; this department specialises in utility projects. This research then analyses and improves the accuracy of the estimated project durations by using data from the utility construction projects.

Based upon related work in the field of project durations, artificial neural networks (ANN) and regression are used to determine the project durations of construction projects. Before the techniques are tested in a case study, the current situation of the contractor is analysed to evaluate the performance of the techniques. Prediction models and scripts are thereby developed based on the two techniques. Different parameters are selected and tested in the model to improve the prediction of project durations of utility projects.

Based on the mean absolute error (MAE), it appears that the ANN and regression techniques do not outperform the current situation with the contractor. The current MAE of the utility construction projects of the contractor is 8.4 days. The MAE of the predicted project duration of the utility construction projects with regression is 9.2 days. The MAE of the predicted project duration with regression increased by 9.5% on average relative to the MAE of the current project duration of the utility construction projects. Furthermore, the MAE of the utility construction projects planned with ANN is 8.52 days. The MAE of the predicted project duration with ANN increased by 1.4% on average relative to the MAE of the current project duration of the utility construction projects. Based on the dataset of the utility construction projects used in this research, the current MAE of the contractor is superior to the MAE corresponding with the ANN technique and the regression technique.

Preface

This report is my Master's thesis, and it's completion marks the final stage of my studies at the Department of Building Environment of the University of Technology in Eindhoven. During this time I can say that I learned, expanded my knowledge and grew as a person working on a relative new topic within the construction industry. The path I took answering my research question about improving the accuracy of project duration with artificial intelligence, turned out to be challenging, sometimes woolly, but always interesting. While researching different prediction tools and useful parameters to predict the accuracy of utility construction projects, I had to dig deeper in the data behind these techniques than I had imagined before I started. Thereby, the data that is needed from Contractor was mostly not available or not useful for this research. However, looking back I am happy that I gained all relevant knowledge, which can help me during my career in the future.

During my graduation time I had support from different people around me. They made efforts to help me further and without them I couldn't reached this final point of my graduation. Firstly, I would like to thank my supervisor Aant van der Zee for the guidance, discussions about the subject and the critical questions during the whole process which helped me to get the best out of my work during the graduation. Thereby, I would like to thank my second supervisor Tao Feng for the support, guidance and the critical questions during the whole process and especially at the end of my graduation period.

Furthermore, I would like to thank the utility development within Contractor for giving me the opportunity to carry out this research. Especially, I would like to thank my external supervisors Jorgen de Kock and Willem van Dijk within Contractor for the guidance, sharing knowledge and critical questions. Thereby, within Contractor I would like to thank also John van der Meer for his practical input and guidance during my graduation process.

Besides, I would like to give my special thanks to my parents for supporting me during my whole student career and giving me the opportunity to pursue a good education. The rest of my family – especially my brothers and sister thank you for always supporting me and being there for me whenever I needed it most.

The last thing I want to say is that I am disappointed about the fact that it was not possible with the script to predict utility construction projects with Artificial Neural Networks, but I really enjoyed writing this thesis. The research made me very ambitious since it covers all the aspects I aspire. Although all, doing this research was a valuable experience I would carry on for the rest of my life.

Executive Summary

In the contemporary building sector there is a higher demand in being more efficient. Due to the decrease of the profit margins in this sector, contractors are more focussed on the cost drivers in the projects in which they are engaged. In addition to the less profitability building market there is also a lack in labour supply and increasing material costs. Therefore the focus to be profitable can be find in having an efficient building process. One of the key elements in having an efficient building process is being more accurate in the time this process will take. This leads in having a good prediction of the project duration. The importance of being able to predict the project duration time is necessary to decide whether a project is profitable or not. Due to the higher competition in this market and the lower profit margins this research will focus on improving the accuracy of project duration with the use of artificial intelligence.

A literature study is conducted to get insights in the different techniques available for predicting project duration. Based upon an overview of the related works in the field of construction time, two techniques are selected to apply in a case study. Machine learning methods as Multi-Linear Regression and Artificial Neural Network are applied in a case study to predict future project durations. The characteristic of the first technique is that its representation is simple and can be easily implemented in daily practice. Neural Networks are used, since literature suggest that they have a major application in forecasting, and are well suited for predictions if there is enough data available.

Therefore, the two artificial techniques known as the Artificial Neural Network (ANN) and Regression techniques are used to analyse the performance of the project durations of Heijmans and compared this to the real performance results of Heijmans. The findings are quite interesting because the current performance is better than the predicted performance by using the artificial techniques. The current mean absolute error of the utility construction projects of the contractor is 8,4 days. The mean absolute error of the predicted project duration of the utility construction projects with regression is 9,2 days. The mean absolute error of the predicted project duration with regression increased on average with 9,5 % compared to the mean absolute error of the current project duration of the utility construction projects. Furthermore, the mean absolute error of the utility construction projects planned with ANN is 8,52 days. The mean absolute error of the predicted project duration with ANN increased on average with 1,4 percent compared to the mean absolute error of the current project duration of the utility construction projects.

From the results it can be concluded that the use of the Artificial Neural Network (ANN) and Regression techniques will not lead to better performance results. The current situation perform much better compared to the results of the regression technique and the ANN technique. The neural network model outperforms the regression technique as expected.

The main limitations of this research is the incomplete nature of the dataset. Therefore, the data of projects within the contractor have not been processed correctly. To conduct this

research, it is important to have a useful dataset to test the created models. This research has opted to use the parameters that have data within the contractor for testing the created models. Another limitation is the usage of limited information. Most of the variables that were selected to test in this research emerged from the availability in the initial dataset.

For future research, several opportunities should be explored. The project duration of utility construction projects currently only relies on knowledge of the planner and some related projects that have been completed in the past. This is insufficient to optimise the project duration within the utility construction industry. A richer dataset combined with an extension of the technique might enhance the already promising results. Furthermore, new prediction techniques such as the ANN and the regression technique, particularly during the process, can enable smaller project durations for utility projects. Incorporation of these techniques can further enhance the technique and increase the precision of project durations, which can yield a competitive advantage for the contractor

Managementsamenvatting

In de hedendaagse bouwsector is er een hogere vraag naar efficiëntie. Door de daling van de winstmarges in de bouwsector, zijn aannemers meer gefocust op de kostenbepalende factoren binnen de projecten die zij uitvoeren. Naast de minder renderende bouwmarkt is er ook een tekort aan arbeidsaanbod en stijgende materiaalkosten. Daarom is de focus binnen de aannemerij om zo winstgevend te zijn binnen een efficiënt bouwproces. Een van de belangrijkste elementen bij het hebben van een efficiënt bouwproces, is meer accuraatheid binnen het bouwproces. Dit leidt in het hebben van een goede voorspelling van de bouwtijd van het project. Het belang van het kunnen voorspellen van de duur van het project is nodig om te beslissen of een project winstgevend is of niet. Vanwege de hogere concurrentie in de bouwindustrie en de lagere winstmarges zal dit onderzoek zich richten op het verbeteren van de nauwkeurigheid van de projectduur met behulp van kunstmatige intelligentie.

Een literatuurstudie wordt uitgevoerd om inzicht te krijgen in de verschillende beschikbare technieken voor het voorspellen van de projectduur. Op basis van een overzicht van de gerelateerde werken op het gebied van bouwtijd, worden twee technieken geselecteerd om toe te passen in een case study. Machine learning methodes zoals Multi-Linear Regression en Artificial Neural Network worden toegepast in een case study om toekomstige projectduren te voorspellen. Het kenmerk van de eerste techniek is dat de weergave ervan eenvoudig is en gemakkelijk kan worden geïmplementeerd in de dagelijkse praktijk. Neurale netwerken worden gebruikt, omdat de literatuur suggereert dat ze een belangrijke toepassing hebben bij het voorspellen en zeer geschikt zijn voor voorspellingen als er voldoende data beschikbaar is.

Daarom worden de twee kunstmatige technieken die bekend staan als de Artificial Neural Network (ANN) en Regressie gebruikt om de prestaties van de bouwtijden van Heijmans te analyseren en te vergelijken met de echte prestatieresultaten van Heijmans. De bevindingen zijn vrij interessant, omdat de huidige prestaties beter zijn dan de voorspelde prestaties met behulp van de kunstmatige technieken. De huidige gemiddelde afwijking van de utiliteitsbouwprojecten van de aannemer is 8,4 dagen. De gemiddelde afwijking van de voorspelde projectduur met regressie van de utiliteitsbouwprojecten is 9,2 dagen. De gemiddelde absolute afwijking van de voorspelde projectduur met regressie steeg gemiddeld met 9,5% in vergelijking met de gemiddelde absolute afwijking van de huidige projectduur van de utiliteitsbouwprojecten. Verder is de gemiddelde absolute afwijking van de geplande utiliteitsprojecten met ANN 8,52 dagen. De gemiddelde absolute afwijking van de voorspelde projectduur met ANN nam gemiddeld toe met 1,4 procent in vergelijking met de gemiddelde absolute afwijking van de huidige projectduur van de utiliteitsbouwprojecten.

Uit de resultaten kan worden geconcludeerd dat het gebruik van de technieken Artificial Neural Network (ANN) en Regressie niet zal leiden tot betere prestatieresultaten. De huidige situatie presteert veel beter in vergelijking met de resultaten van de regressietechniek en de ANN-techniek. Het neurale netwerkmodel presteert zoals verwacht beter dan de regressietechniek.

De belangrijkste tekortkomingen van dit onderzoek is dat de dataset niet volledig was. Daarom zijn de gegevens van projecten die door de aannemer zijn uitgevoerd, niet correct verwerkt in dit onderzoek. Om dit onderzoek uit te voeren, is het belangrijk om een bruikbare dataset te hebben om de gemaakte modellen te testen. In dit onderzoek is alleen gekozen voor parameters waarvan de aannemer datasets van heeft.

Voor toekomstig onderzoek moeten verschillende mogelijkheden worden onderzocht. De projectduur van utiliteitsbouwprojecten is momenteel alleen afhankelijk van kennis van de planner en enkele gerelateerde projecten die in het verleden zijn voltooid. Dit is onvoldoende om de bouwtijd binnen de utiliteitsbouw te optimaliseren. Een rijkere dataset in combinatie met een uitbreiding van de techniek zou de al veelbelovende resultaten kunnen verbeteren. Bovendien kunnen nieuwe voorspellingstechnieken zoals de ANN- en de regressietechniek, in het bijzonder tijdens het proces, de duur van bouwprojecten verbeteren. De integratie van deze technieken kan de techniek verder verbeteren en de daadwerkelijke duur van projecten verhogen, wat een concurrentievoordeel voor de aannemer kan opleveren.

Table of Contents

List of Figures

List of Tables

List of Acronyms

AEC	Architecture, Engineering, and Construction
ANN	Artificial Neural Network
ANOVA	Analysis of variance
CRISP-DM	Cross-Industry Standard Process for Data Mining (Crisp-DM)
DSRP	Design Science Research Process
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MIP	Mixed Integer Programming
MLP	Multi-Layer Perceptron
NN	Neural Networks
SPSS	Statistical Package for the Social Sciences
UCI	Utility Construction Industry
UCP	Utility Construction Projects

1. Introduction

This chapter describes the context of this research project. It first provides the problem definition for this research and presents the research questions of this project. It then discusses the research goals and research design of this project. Finally, it briefly introduces the case study company and presents the thesis structure.

1.1. Problem definition

The construction industry in the Netherlands contributes substantially to the working environment and quality of life. The construction industry in the Netherlands consists of three types of projects, namely civil engineering, residential, and utility projects. Table 1 provides an overview of the turnover associated with the three types of projects (Feiten & Cijfers). The turnover related to residential projects is the highest at 43%, followed by that related to utility projects, which is 31%. The turnover related to civil engineering projects is the lowest at 26%. The total turnover of the construction industry in 2016 was 58.5 billion euros.

Turnover of the construction industry in 2016 (euros)	
Civil engineering projects	15.33 billion euros
Utility projects	18.08 billion euros
Residential projects	25.08 billion euros

Table 1 - Turnover of the construction industry in the Netherlands ((Feiten & Cijfers, 2016))

This research primarily focusses on schedule delays in the planning phase of utility projects. The utility projects are mostly recurring projects with some differing features, such as simple environments, high complexity, dynamic project organisation, and long-term duration (Zou, Zhang & Wang, 2007). Within the utility construction industry, it is possible to apply repetition and reuse knowledge because projects are generally highly similar (Abbasi et al., 2018). Uncertainties are also caused by the environment and climate during utility projects (Al-Zarrad, 2017). These uncertainties are important to manage before the execution phase is initiated. The tools that can be used to manage these uncertainties are highly contingent upon expert multi-disciplinary knowledge and experiences (Zou, Kiviniemi & W. Jones;, 2017). Multi-disciplinary knowledge and experiences provide a knowledgeable perspective of the uncertainties for each involved individual who examines the process of a utility construction project (Gorse, Stafford, Miles-Shenton, Johnston, Sutton & Farmer, 2012). According Zou et al. (2017), it is nevertheless important to quantify all potential and relevant uncertainties of projects. Therefore, it is also important to manage the valuable cognition effectively.

Construction planning is often characterised by significant uncertainties due to a lack of information and because many activities are performed by various sub-contractors (Thomas NG & Tang, 2010). Construction planning is a fundamental and challenging activity in the management and execution of construction projects (Zaitseva & Rogovenko , 2017). It involves choices regarding technology, defining work tasks, estimating required resources and durations for individual tasks, and identifying interactions among the different work tasks.

Many projects are currently planned improperly in the building industry (Doree, Holmen, Caerteling, 2003). To improve the prediction of project durations, a technique should be implemented to determine whether this could be improved.

1.2. Research question(s)

This project aims to improve the prediction of project durations in construction management. Using various parameters in the technique can enable the development of different models to predict future project durations. Improving these predictions could save millions of euros in the building industry.

The main research question can be derived from this research objective:

> ***'How can accurate and reliable historic project durations be used to predict a more efficient construction project duration?'***

To answer the main research question, the following sub-research questions have been formulated:

[1] *Which methods are currently available for predicting project duration within the utility construction industry?*

[2] *How can the prediction of project duration in the utility construction industry be made more accurate?*

[3] *What parameters could influence the prediction of project durations?*

[4] *How does the proposed method perform in a case study?*
- What is the current performance?
- What is the performance of the imposed method?

1.3. Research design

The research design is intended to address the research questions mentioned in paragraph 1.2. This research approach is based on the DSRP (Distinctions, Systems, Relationships, and Perspectives) method, which is used in the demonstration phase of the data analysis. This method consists of two premises. The first premise holds that humans build knowledge and that knowledge and thinking are in a continuous feedback loop. The second premise holds that knowledge changes (Senge, 2017).

1.3.1. DSRP method

The DSRP method consists of six steps, which are presented in Figure 1. Firstly, the specific research area is analysed, and the definition of the problem is provided. A justification of the value of the solution is also presented. The second step involves stating the objectives of a solution. Therefore, the research questions are specified to extract the main objectives.

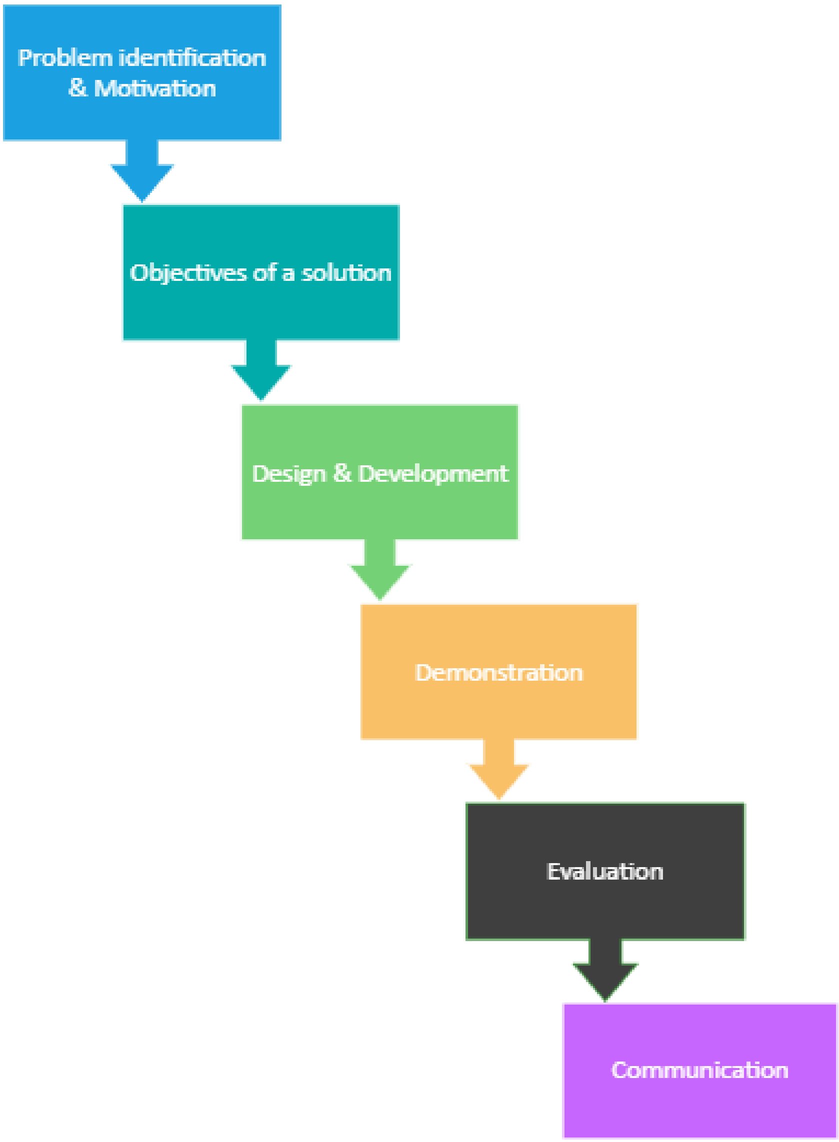

Figure 1 - The DSRP method (Peffers, Tuunanen, Rothenberger & Chatterjee, 2006)

The Design and Development component of the DSRP focusses on a literature study to review the related parameters available for predicting project durations and the techniques used. After providing clear insights from the current research, this research paper proposes a method that combines the optimal elements of past research; this research specifically applies ANN. In addition, it analyses the available project data in construction management, which addresses the research questions [1].

In the demonstration process, the proposed method is tested using the model and software. In this phase, the performance of the technique is measured using a real dataset. To accomplish this, the CRISP-DM framework is used as a guide, as illustrated in Figure 2. The modelling component of the CRISP-DM provides insights into how to predict project durations; this therefore addresses the research question [2].

Thereafter, the proposed method is evaluated through a comparison of the generated project durations with the actual durations. The research question [3] is addressed by making the predictions more accurate using different layer sizes and nodes in the ANN.

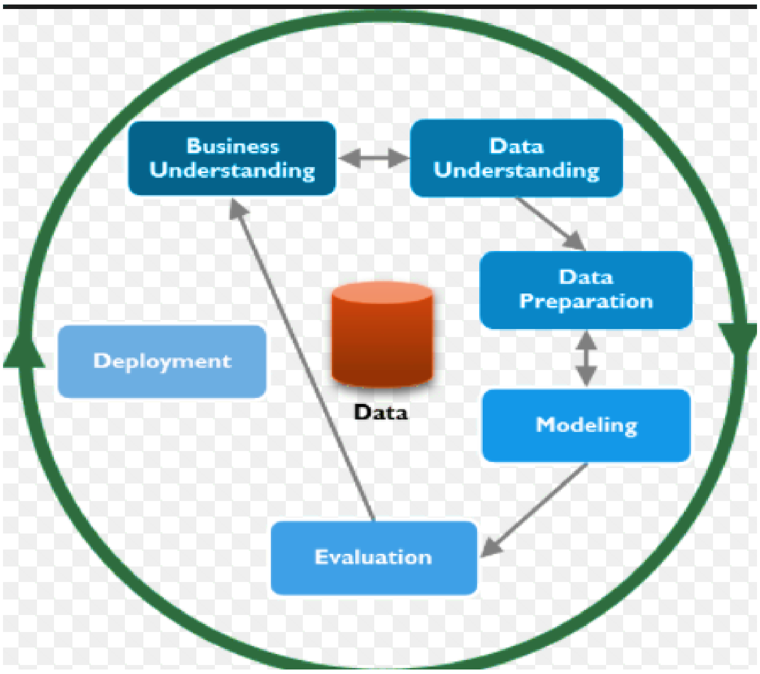

Figure 2 - The CRISP-DM (Peffers, Tuunanen, Rothenberger & Chatterjee, 2007)

Finally, feedback about the overall research and practical implications are provided in the communication phase of the DSRP, which also addresses the research question [4].

1.3.2. Conceptual research framework

The conceptual research framework is intended to elaborate upon the Design Science Research Process (DSRP) according to Peffers et al. (2016). This conceptual research framework helps to enable high-quality research in the utility construction industry (UCI). This conceptual research framework is combined with the CRISP-DM framework, which is used in the demonstration phase for data analysis. Figure 3 illustrates the conceptual research framework, which represents the design of the remainder of this thesis. It consists of three phases, including the DSRP phase, research phase, and deliverables.

Firstly, the DSRP phase consists of Design and Development, which is subdivided in the literature review and the description of techniques and the modelling approach. The literature review consists of two deliverables. The deliverables are the quantitative factors which influence construction projects and prediction tools in the construction industry. The description of the techniques and modelling approaches consists of three deliverables, including the explanation of the modelling approach, parameters for modelling, and the evaluation method.

Secondly, the DSRP phase consists of the demonstration phase, which is subdivided into created tools and a case study. The created tools are subdivided into a tools and script description and the characteristics of the tools and script. The case study is subdivided into data preparation, data cleaning, current performance, and the performance of the two techniques applied.

Thirdly, the DSRP phase consists of an evaluation phase, which is subdivided into the conclusion and recommendations. The conclusion is subdivided into a comparison of the two used prediction techniques and factors which influence the utility construction projects. The recommendations are subdivided into recommendations for future research and recommendations for the contractor.

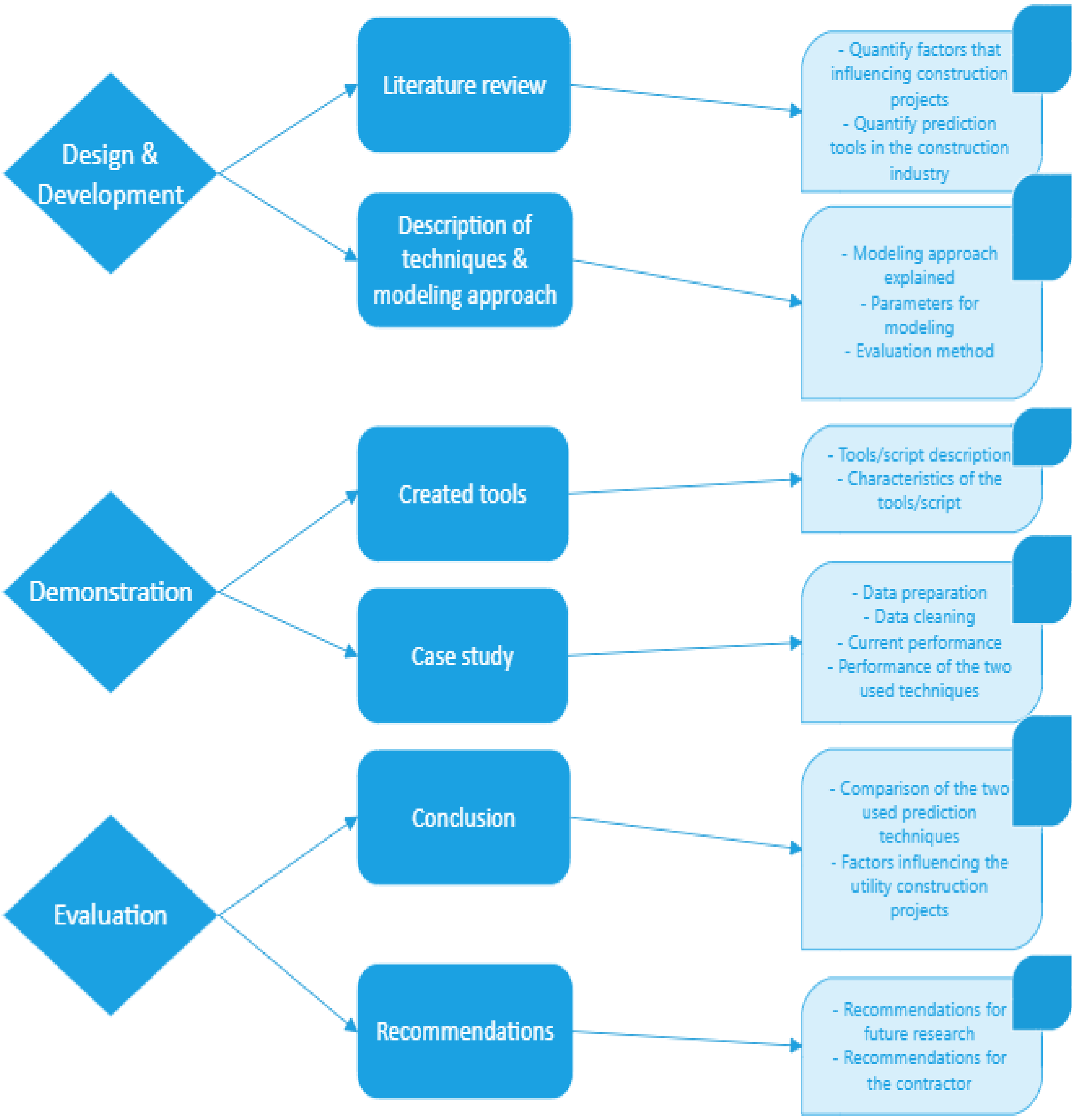

Figure 3 - Conceptual research framework

1.4. The scientific and practical importance of the thesis

The graduation research has been elaborated for the Master of Construction, Management, and Engineering. This master's thesis focusses on the urban environment and also combines entrepreneurship and industrial engineering elements. To expand upon this research, research into the prediction and delays of construction projects is conducted. In the Netherlands, no research has investigated the factors that cause delays and affect the prediction phase of utility projects. This research aims to develop greater insights into the prediction phase of utility construction projects and to delineate the factors which influence the delays. Construction company contractors have participated in the execution of this research. Contractors are active in the construction industry and are therefore subject to the prediction of utility construction projects. Contractors in the Netherlands predict construction projects with Lean Construction. Lean Construction is an approach to building that is iterative. Planning, design and building are done in smaller chunks one at time with a detailed approach called "the Last Planner Method'', as shown in figure 4.

Figure 4 - "the Last Planner Method'' (parsonscorp, 2014)

Lean planning in construction is based on the commitment to deliver a certain element of work and element of time per day starting from the end and moving back to the beginning lean planning is all about collaboration. All stakeholders collaborate and plan together; owner, contractor, sub-contractors, designers, engineers, schedulers, cost engineers, etc. The disadvantage of lean is that if a party is not ready on time, the planning is no longer correct and needs to be changed. Despite the fact that there is planned with lean planning the last years, there are still many projects planned improperly in the building industry. To improve the prediction of project durations, a technique should be implemented to determine whether this could be improved.

1.5. Scope and requirements

The scope has been determined based upon data availability and in collaboration with the construction company.

- Utility construction data must be available and useful for predicting utility project durations with software;

- To achieve a fair representation of the project duration, only available data with the same parameters has been selected;

- There is not sufficient useful utility data about the surfaces of window frames, doors, and walls. Therefore, this research focusses on the surfaces of buildings in the context of predicting the accuracy of utility construction projects.

1.6. Reading guide

Chapter 2 presents a literature review of the factors which influence the project duration of utility construction projects and the existing techniques for predicting the utility construction projects. Chapter 3 discusses the relevant parameters, the two selected techniques, and their modelling approach. Therefore, chapter 3 also develops an ANN model for predicting the duration of utility construction projects.

Furthermore, chapter 4 consists of a case study to validate the proposed techniques. It performs an evaluation by comparing the performance of the proposed method with historic data. In addition, chapter 4 explains the characteristics of the developed tool in Matlab. After all research steps have been delineated, chapter 5 presents an overall conclusion and potential applications for the case company. Chapter 5 also elucidates the limitations of the selected method and all recommendations for future research.

2. Literature review

This chapter analyses literature from several sources to determine the factors that influence construction projects. This research elucidates relevant prior literature, which provides the theoretical background. Firstly, this chapter describes the general factors that influence the project duration of construction projects. Secondly, it then delineates the factors that influence the project duration of utility construction projects. Furthermore, the third section analyses the prediction tools. The final section concludes the literature review.

2.1. Factors which influence construction projects

Construction projects are generally disturbed by poor performance factors such as cost overrun, delays, construction faults, and compromised quality (Zairoh et al., 2017). Construction projects often experience delays and squandered time due to various factors; they therefore suffer from unfavourable consequences (Zidane et al., 2018). Delays are common in the scheduling of construction projects and can cause considerable losses to project parties (Abbasi et al., 2018). Several articles have analysed the factors that affect construction projects. Most articles have prioritised identifying the factors that cause delays and provided insight into the perspectives of different parties, namely clients, contractors, and consultants.

2.1.1. General construction projects

The analysed articles delineate various factors that influence construction projects. Kikwasi (2013) and Gunduz (2016) describe over 10 such factors, including poor communication, coordination, and risk management during the project. Additionally, the most pivotal factors described by Gluszak and Lesniak (2016), Marzouk et al. (2014), and Gardezi et al. (2013) include incompetent project teams and ineffective delay penalties. Senouci et al. (2016) contend that an incompetent project team is the most crucial factor that influences construction projects. An incompetent project team may result in many mistakes, which renders it difficult to complete the project successfully.

Zaitseva et al. (2018) and Kikwasi (2013) note that many aspects may influence the delay of the project. They assert that a ‘shortage of materials and equipment in the local market’, ‘delays in approving permits’ and ‘project risks’ are the most important factors that influence construction projects. Conversely, Zidane and Anderson (2018) argues that, among other factors, the late delivery of materials and equipment, unrealistic tender cost estimation, and poor contract administration are the most significant factors in influencing construction projects. Furthermore, Adam et al. (2014) and Arditi et al. (2006) note that the factors that exert the greatest influence on construction projects include discrepancies in the contract documents. In addition, Bhargava et al. (2010) and Gurgun et al. (2013) hold that aside from these discrepancies, design changes during the construction projects can cause the construction projects to be delayed. Shehu et al. (2014) and Abas et al. (2018) contend that a lack of information exchange between parties is also a pivotal factor because it can delay the process of the construction project. P.Gonzalez et al. (2014) note that the factors that influence construction projects occur in the contract phase; they assert that unrealistic enforced contract duration and poor contract management are important factors in influencing construction projects.

In the reviewed articles, the factors were classified into groups based on frequency. The quantification of the reliance of one factor over other factors was addressed in multiple articles. These articles focus on the relationship between the factors that influence construction projects and delays in construction projects.

2.2. Factors which influence project duration of utility construction projects

The number of factors that influence the project duration of utility construction projects differ in each article, as indicated in Appendix 2. Appendix 1 presents the reviewed articles and the authors. These articles are used in appendix 2 to provide an overview of the factors that influence utility construction projects. It lists 17 main factors and over 30 sub-factors that influence construction projects. In the articles analysed, some factors are mentioned more frequently than others. This does not mean that these factors contribute more. During the analysis of the literature, it is important to consider each article and which cause can be assigned to each factor without repeating the factors.

2.2.1. Utility construction projects

Several articles address factors the many factors that influence utility construction projects. For example, Al Jurf et al. (2010), Abbasi et al. (2018), and Bhargava et al. (2010) assert that a shortage of materials and equipment in the local market is an important factor. Furthermore, Abas et al. (2018) and Gluszak et al. (2015) contend that delays in approving building permits influence utility construction projects in the definition phase. This influence affects the project duration at the beginning of a project. For example, Rajguru (2015) discusses the factors 'shortage of materials and equipment in the local market' and 'too expensive materials and equipment in the local market' as two separate factors, while other articles combine these factors into one factor.

Additionally, P.Gonzalez et al. (2014) describe the factor 'lack of team alignment and conflicts management' as one accumulated factor, while in other articles refer to it these factors separately as 'lack of team' and 'conflict management'. Larsen et al. (2015) mention the factor 'unrealistic tender cost estimation', while other articles refer to a combined similar factor 'unrealistic tender cost and contract duration estimation'. The article by P.Gonzalez et al. (2014) frequently mentions the most main factors. It describes thirteen main factors, including the following: 'delays in approving permits', 'late delivery of materials and equipment', 'delays in monthly payments from owner', 'incompetent project team', 'design related issues', 'lack of team alignment and conflicts management', 'changed conditions of the project', 'escalation of basic construction material prices in local market', 'late delivery of materials and equipment', 'general labour shortages', 'low labour productivity', 'unrealistic tender cost estimation', 'unrealistic contract duration estimate', and 'weather conditions'.

The three most commonly mentioned factors that influence utility construction projects include 'late delivery of materials and equipment', that is mentioned 18 times, 'weather conditions', which is mentioned 14 times and 'project risks', which is also mentioned 14 times.

Table 2 indicates the most commonly mentioned factors that influence utility construction projects based on an analysis of the literature. The factor 'delays in approving permits' is mentioned in 10 articles. This factor strongly influences construction projects, according to articles by Zaitseva et al. (2017) and Gluszak et al. (2015). 'Delays in approving permits' exerts a limited influence on construction projects, according to articles by Al Jurf et al. (2010), Shehu et al. (2017), and Gunduz (2016). Additionally, the factor 'delays in conducting inspections and testing of work' is referred to in seven articles. The articles by Abbasi et al. (2018) and Arditi et al. (2006) described this factor as a minimally influencing factor in construction projects. The articles do not provide information about whether this factor exerts a strong influence on construction projects in contrast with other factors. Furthermore, the factor 'incompetent project team' is mentioned in 11 articles. This factor exerts a strong influence on construction projects, according to the articles by Bhargava et al. (2010) and Gluszak et al. (2015). The articles by Kikwasi (2013) and Al Jurf et al. (2010) note that this factor exerts a limited influence on construction projects.

Ranking factors influencing the projects		
Ranking	Factor influencing the projects	Frequency
1	'Late delivery of materials and equipment'	18
2	'Weather conditions'	14
	'Project risks'	14
3	'Changed conditions of the project'	13
4	'Incompetent project team'	11
5	'Lack of team alignment and conflict management'	10
	'Delays in approving permits'	10
6	'Construction methods'	9
	'Design-related issues'	9
	'Changed conditions of the project'	9
7	'Shortage of materials and equipment in the local market'	8
	'Escalation of basic construction material prices in local market'	8
	'General labour shortages'	8
	'Unrealistic tender cost estimation'	8
	'Unrealistic contract duration estimate'	8
8	'Poor contract administration'	7
9	'Delay in monthly payments from owner'	5

Table 2 - Ranking of factors which influence the projects

The factor ‘lack of team alignment and conflicts management’ is referred to in 13 articles. They state that the influence of this factor on construction projects is highly different. Some articles assert that this factor exerts a limited influence on the construction industry, and some articles indicate that this factor exerts a strong influence on the construction industry. The articles by Al-Zarrad (2017), Bhargava et al. (2010), and Marzouk et al. (2014) also contend that this factor exerts a strong influence on the construction industry. The other articles provide no information about whether this factor exerts a limited influence on the construction projects in contrast with other factors.

As indicated in Table 2, the analysed articles also describe the three factors mentioned with the lowest frequency. For example, five factors are only mentioned eight times, including ‘shortage of materials and equipment in the local market’, ‘escalation of basic construction material prices in local market’, ‘general labour shortages’, ‘unrealistic tender cost estimation’, and ‘unrealistic contract duration estimate’. The factor ‘poor contract administration’ is mentioned seven times, and the factor ‘delay in monthly payments from owner’ is mentioned five times.

2.3. Tools for predicting utility project duration

This section presents a cross-analysis of related works concerning the field of techniques for prediction tools. This section aims to provide an overview of the techniques and methods used in past research to predict the accuracy of utility projects. The related works have been extracted by analysing different articles. It provides an overview of the available literature concerning predicting project duration within the utility construction industry.

2.3.1. Predicting tools

The tabular model created with the analysis of the articles in Appendix 3 explains the content of the articles. The first column indicates the articles. The other columns assess all of the analysed articles in terms of the dimensions listed above. This research focusses on the accuracy of the project duration within the utility construction industry. Therefore, the scope of this research encompasses accuracy in predicting utility construction projects. Most research papers from the literature review (appendix 3) are discussed here. Based on these descriptions, this section ultimately selects a technique. Appendix 3 presents the overall results of the related work analysis.

Said et al. (2013) applied two methods of managing demand variability in seasonal time series using artificial neural networks. According to this article about forecasting methods, neural networks models are properly configured. Zhang et al. (1998) addresses the application of ANNs in forecasting. This article aimed to provide a synthesis of academic research concerning predicting and providing a comprehensive overview of ANN modelling issues. El Hassani et al. (2013) contend that demand forecasting is pivotal in the supply chain of today's companies. Because of predicting methods, neural networks models can deliver optimal results if they are properly configured.

El Hassani et al. (2013) applied two approaches based on multilayer perceptron which have been developed to predict demand. Based on a judicious selection of the architecture and parameters of the neural network, both approaches have yielded positive results. Nikzad et al. (2015) performed a comparative study between a regression analysis and a multilayer feed-forward neural network. This was conducted to estimate their abilities to predict drying ratios. Nikzad et al. (2015) subsequently noted that the ANN model describes the drying behaviour more accurately.

Samanta et al. (2001) indicate that an ANN-based procedure was presented for fault diagnosis of rolling element bearings using features extracted directly from time-domain vibration signal segments through simple processing. Samanta et al. (2001) state that the artificial neural networks consist of four or five input nodes, two hidden layers with 16 and 10 neurons, respectively, and two output neurons. Furthermore, the success rate for training was almost 100%, and that of the test was approximately 98% to 100%. Ghiassi et al. (2005) present a dynamic artificial neural network model for forecasting time series events. This model was created to offer an alternative architecture to existing artificial neural networks. Ghiassi et al. (2005) mention that a group of benchmark time series events were compared to demonstrate the effectiveness of this model.

2.3.2. Comparison between techniques

As indicated in Appendix 3, Said et al. (2013) yielded the same results in the evaluation as Khosravi et al. (2015). Said et al. (2013) focusses on project duration in general, and Khosrava et al. (2015) focusses on specific logistic projects. Therefore, this article is excluded in this research. Furthermore, the article by Samantha et al. (2001) described the performance of the techniques that are used in the article as well as the implementation in practice. In addition, the article by Patel et al. (2016) discussed merely the performance of the techniques used in the article. These articles indicate that the techniques are useful for practical implications in this research.

The article by Remus et al. (1994) and the article by Kohzadi et al. (2007) applied a causal model and a statistical ANN. The articles developed a causal model that is an abstract model which describes the causal mechanisms of a system. This model must express more than correlation. These articles aim to present forecasting with artificial neural networks. The article by Kohzadi et al. (2007) also applied a regression; it is important to determine how it can be implemented in this research. Since the literature selection focusses on the prediction of project duration within the utility construction projects, implementing artificial neural networks in a case study may be interesting.

The article by Said et al. (2013) proposed a clear technique for predicting the accuracy of a project because the article used the statistical ANN. This article used a historical dataset to predict project durations. According to Said et al. (2013), neural network models are properly configured. As indicated in appendix 3, the aspects and prediction techniques used in the articles have been analysed. The literature review provides a clear overview of the techniques and methods used in the articles. This research focusses on two specific prediction techniques, namely regression and artificial neural networks. These techniques make predictions by learning from datasets.

2.4. Modelling approach to predicting project duration

This section delineates the modelling approach to predicting the accuracy of utility construction projects. The techniques used include the regression and artificial neural networks.

2.4.1. Selection of methods and techniques

As we look back to the problem definition, we need a reliability prediction method that can handle factors that influence the project duration. From literature there are several methods. Three of them are listed below:

1. Multiple Regression Analysis;
2. Artificial Neural Networks;

A multiple regression analysis is conducted to look for any relationships which could lead to a good project duration. Models like regression analysis can indicate simple linear relations. These statistical analyses models can be used to estimate project durations in the utility construction industry. Hair et al. (2005) shows that multiple regression analysis can be performed, since the situation of this case study is: One dependent variable (metric) =six independent variables (metric & non metric).

Neural network seemed also a good choice to predict the project duration since this method can use the factors and its data that are identified in the case study as input parameters to predict reliability. This is possible because neural networks can handle the properties of the found factors like nonlinearity. Data found in the study case could change or update very often because of new found data sources available. This is no problem for the neural network method since adapting the model and retraining can be done very fast. This is also an advantage when new data becomes available in the near future at contractors. When a factor is put into the NN which does not have any influence on the project duration, statistics will teIl, and the network will automatically ignore the certain factor. But the opposite is also possible; a neural network can make a relationship between two variables based on data while in a real situation the independent variable does not change the dependent variable.

In this research, the ANN and regression technique are used on the basis of the selection in the literature review. For the selection of methods and techniques that are used in this research, the articles are assessed on the following dimensions:

1. Requirements
2. Prediction technique used
3. Performance of the technique and evaluation
4. Implementation in practice

A distinction is made when considering the methods and techniques that are used in the articles. The focus in this research will be on two specific predicting techniques; Artificial Neural Network and Regression. These techniques aim at making predictions for any target variable by learning from a dataset. Regression is an attractive model because the representation is simple and can be easily implemented in daily practice. ANNs are used for the fact that they have a major application in forecasting (Sharda, 1994), and are well suited for predictions if there is enough data available. Since this research will consider a case study implementation, it would be an advantage if the proposed method or technique is already implemented in practice. In addition, the performance of the technique in combination with the experiment used for evaluation has also influence on selecting the initial method or technique to use for this research. This implies that articles that achieve a low performance after implementing their technique will be neglected.

2.4.2. Regression

Multiple regression analysis is the simplest prediction method. It uses the data to create a function which ideally fits the output data (Hair et al. 2005). The regression analysis identifies the relationships between all of the independent variables. This approach is the most commonly used analysis in the statistical industry. To continue the research with a regression, parameters are used from the literature review to develop an estimated regression for predicting the project duration within the utility construction project.

This research uses the multi-linear regression. This regression model is the first model developed for application in the case study. The multi-linear regression model indicates the relationship between the variable used for prediction and the variable mentioned as a target. The aim of the regression technique in this research is to improve the accuracy of the prediction of utility construction projects. The model is implemented in a case study of the contractor.

Important assumptions follow the regression analysis. A high positive value indicates a positive linear relationship, and a high negative value indicates a negative linear relationships. Additionally, a correlation coefficient of 0 indicates no linear relationship between two variables. This research uses the same cut-off value as Jeong and Rilett (2004), namely 0.15. Models such as regression analysis can be used to indicate simple linear relationships or estimate reliability. Hair et al. (2005) state that multiple regression analysis can be performed, since the situation of this case study is as follows: one dependent variable (metric) = six independent variables (metric and non-metric).

The sample size affects the possibility of generalising the result of a regression model. Hair et al. (2005) propose the general rule that the ratio of observations to independent variables should never be lower than 5:1. However, this is the minimum ratio; 15 or 20:1 is preferable. Since there are independent variables, this ratio is easily satisfied.

2.4.3. Artificial neural networks

Artificial neural networks are used as a calculational tool for analysing the data and creating models that help to identify patterns and structures in the dataset. This research classifies the available dataset into three samples, namely training, testing, and validating samples. The samples are divided as follows: 70% training data, 15% testing data, and 15% validating data. The next step in this research involves creating the model in Matlab.

The artificial neural networks consist of three layers. The input layer consists of different parameters and is highly important because selecting the proper parameters results in higher performance. To determine the relevant parameters to use in the model in Matlab, a feature selection is necessary. This entails selecting a subset of relevant parameters that time planners use in the utility construction industry to create the model. The model is then tested by using one parameter for 100 iterations.

After the iteration process in Matlab, the ideal parameter is selected, and the model is tested again. The ideal parameter is then used as an initial starting point in combination with the parameters that exert a more limited influence on the project duration. This iterative process is repeated until all parameters are included in the model in Matlab. The results of the different iterations are then visualised in a graph. The subset of the parameters with the highest performance is then used as the initial starting point, what means that the ideal parameter is the input layer.

After the input layer is selected, then it is important to determine the hidden layers. Zhang et al. (1998) contend that using more than one layer offers more benefits than merely one layer. In this research, the model that is developed has two hidden layers. According to Zhang et al. (1998), the research performed in this article provides different backbones that are useful for determining the optimal number of neurons in each layer. Each hidden layer consists of an optimal number which is established during the run of the model in Matlab.

This research uses two hidden layers. This means that different combinations of neurons are applicable in the model. To determine the optimal combinations, different solutions have been tested with the model. To yield a representative output, the solutions run 10 iterations. After the number of hidden layers and the optimal solution are known, the network can be run on the validation dataset. The validation dataset can yield a representative output.

2.5. Conclusion

This literature analysis aims to determine the factors that exert a strong influence on utility construction projects. Furthermore, this literature review is necessary to analyse the prediction techniques of utility construction projects which are useful for improving the accuracy of predictions of project duration for utility construction projects.

The results are varied. The comparison of the articles demonstrates that the most important factors that influence the utility construction industry relate to the process. These factors include late delivery of materials and equipment, weather conditions, and project risks. According to the literature review, the late delivery of materials and equipment exerts the strongest influence on utility construction projects. Secondly, bad weather conditions exert a negative effect on the process of utility construction projects. Thirdly, the literature review indicates that the delay of projects is predictable based on project risks. Project risks are classified into three levels, including high, middle, and low.

This chapter has aimed to provide an overview of the various techniques that are currently available for predicting utility construction projects by conducting a literature study. In conclusion, the literature review has demonstrated that the following parameters are highly important. Firstly, the most important parameters include the following: materials and equipment in local markets, amount of materials and equipment, date of approving permits, delivery of materials and equipment, inspections and testing moments during execution, date of payments from the owner, project risks of utility construction projects, construction methods of utility construction projects, members of project team. and design errors. The other important parameters are as follows: project risks of utility construction projects, construction material prices, amount of workers, productivity of workers, surface of m2, estimated project duration, season start project, and contract administration.

The literature has highlighted many techniques that are highly useful for estimating project durations in the construction industry. The most important techniques and the techniques used during this research include artificial neural networks and multi-linear regression.

3. Techniques for predicting the duration of utility construction projects

This chapter explains a component of the Design and Development phase of the DSRP as well as the proposed techniques used in this research. It translates the factors that influence predicted project duration into parameters. These parameters are necessary to predict the accuracy of the utility projects of the contractor. Not all of the parameters can be used for prediction because some datasets of the parameters are not available within the contractor. Secondly, a multilinear regression model has been created to predict the duration. This regression technique combines different parameters that may influence the prediction of the project duration of utility construction projects.

Thirdly, an ANN model has also been developed to predict the accuracy of utility construction projects. The attributes necessary for developing the model are thereby presented and specified. Finally, the relevant prediction variables are defined to predict the accuracy of utility construction projects.

3.1. Relevant parameters

Predictions of actual project duration can only be initiated when the parameters are determined. Chapter 2 has determined the factors which influence utility construction projects based on the literature review. This section transforms the factors that influence utility projects into parameters that are useful for predicting the accuracy of utility construction projects. Table 3 indicates which factors are changed into parameters and are used in this research.

Table 3 illustrates that the parameters used in this research to estimate the accuracy of the utility construction projects including the following: 1) type of utility building, 2) estimated project duration, 3) season start project, 4) surface m2 and 5) project risks.

The first parameter, ‘type of utility building’, consists of five classes. These classes are as follows: 1) office building, 2) hospitals or care centres, and 3) unique projects. The second parameter, ‘the estimate project duration’, is not subdivided. This parameter could not be subdivided because this is a prediction of a project was performed by the contractor itself. The third parameter, ‘season start project’, indicates the season when a project is initiated and is subdivided into four classes. These classes are as follows: 1) autumn, 2) winter, 3) spring, and 4) summer.

The fourth parameter, ‘floor surface m2’, indicates the size of the project. This parameter is subdivided as follows: 1) less than 2000 m2, 2) between 2000 m2 and 5000 m2, 3) between 5001 m2 and 10000 m2, 4) between 10001 m2 and 15000 m2, and 5) 15000 m2 or more. The project risks are defined with the knowledge of the location of a project. This parameter is subdivided as follows: 1) high risks, 2) middle risks, and 3) low risks.

Factors that influence utility construction projects	Parameter created	Relevant for prediction of utility construction projects in ANN	Dataset available within contractor	Useful parameter for this research
'Shortage of materials and equipment in the local market'	Amount of materials and equipment	Yes	No	No
'Delays in approving permits'	Date of approving permits	No	No	No
'Late delivery of materials and equipment'	Delivery of materials and equipment	Yes	No	No
'Delays in conducting inspections and testing of work'	Inspections and testing moments during execution	Yes	No	No
'Delay in monthly payments from owner'	Date of payments from owner	No	No	No
'Project risks'	Project risks of utility construction projects	Yes	Yes	Yes
'Construction methods'	Construction methods of utility construction projects	Yes	No	No
'Incompetent project team'	Members of project team	Yes	No	No
'Design related issues'	Design errors	No	No	No
'Lack of team alignment and conflicts management'	-	No	No	No
'Changed conditions of the project'	Project risks of utility construction projects	Yes	Yes	Yes
'Escalation of basic construction material prices in local market'	Construction material prices	No	No	No
'General labour shortages'	Amount of workers	Yes	No	No
'Low labour productivity'	Productivity of workers	No	No	No
	Surface of m2	Yes	Yes	Yes
'Unrealistic tender cost estimation'	Tender cost estimation	No	Yes	No
'Unrealistic contract duration estimate'	Estimate project duration	Yes	Yes	Yes
'Weather Conditions'	Season start project	Yes	Yes	Yes
'Poor contract administration'	Contract administration	No	No	Yes

Table 3 - Useful parameter for this research

Table 4 presents an overall schematic overview of the parameters and the whole dataset that is necessary to predict the accuracy of the utility construction projects. This table provides a summarised overview of the data preparation process and is the supporting schema for this chapter.

Variable	Specification
Type utility buildings	Type of project and which buildings are created: 1. Office building; 2. Hospitals or care centres; 3. Unique projects.
Estimate project duration	The estimated duration between the start time and the ending time of the utility construction project.
Season start project	1. Autumn; 2. Winter; 3. Spring; 4. Summer.
Floor surface m2	1. Less than 2000 m2; 2. Between 2000 m2 and 5000 m2; 3. Between 5001 m2 and 10000 m2; 4. Between 10001 m2 and 15000 m2; 5. 15000 m2 and more.
Project risks	1. High risks; 2. Moderate risks; 3. Low risks.

Table 4 – specification of the parameters

3.2. Modelling approach

This section elaborates upon the approach of the models selected in chapter 2 based on the literature review. It expounds upon the ANN and then the regression technique. It also explains the key constructs of the two techniques applied. Firstly, it also develops the ANN model with input from Jeong and Rilett (2004) and the state-of-the-art analysis by Zhang et al. (1998). Secondly, it develops a multi-linear regression model, as in the paper by van Duin et al. (2015) and Jeong and Rilett (2004). Finally, it discusses an evaluation of the techniques.

3.2.1 Background of artificial neural networks

The term 'neural network' is derived from the field of biology. It describes how the human brain is organised. The human brain consists of an immense number of elementary cells called neurons. All of these elementary neurons are highly interconnected and therefore form a complex network, called a neural network.

These neurons are highly simple elements. They consist of a soma, the kernel, and an outgoing tentacle attached to the soma, as well as an axon and incoming tentacles called dendrites. The dendrites of each neuron are connected with the axons of other neurons. These connections are formed with the support of elements called synapses. Through the synapses, the dendrites of a particular neuron measure the activity of the axon to which they are connected and propagate this activity forward to the soma of the neuron from which they originate. This effort may be either excitatory or inhibitory, and the measure of influence of one particular axon activity is determined by the synapses that connect to that axon, as illustrated in figure 5. All activity has been demonstrated to be some form of electrochemical action. At the soma, a summation takes place, which means that the action that arrives at the soma from the dendrites is integrated into the total neural activity level. The brain possesses as many as 10^{10} neurons, and every neuron connects to up to 10,000 other neurons, resulting in over 10^{14} connections.

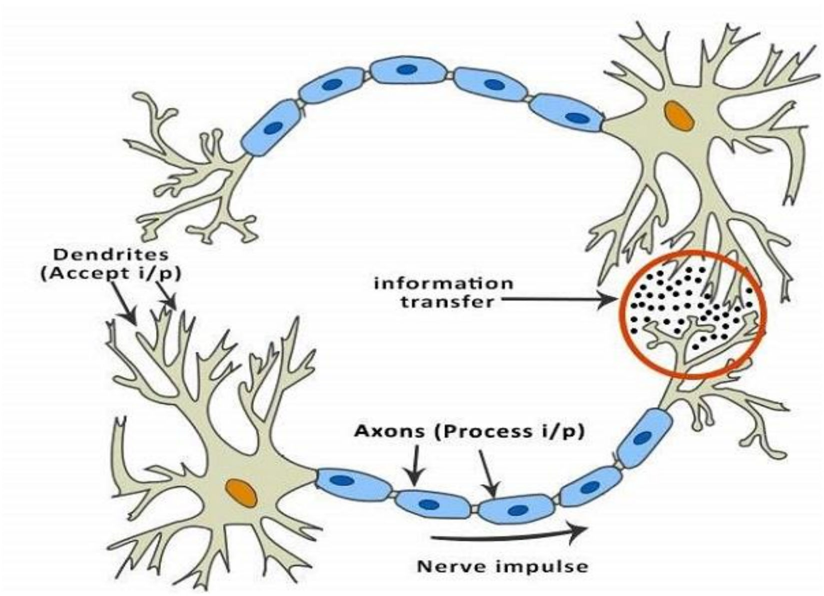

Figure 5 - Neuron (Tutorialspoint, 2018)

ANNs were originally developed to mimic the human brain. They consist of a neural system composed of numerous interconnected processing elements called neurons. All of these neurons in a network can collectively perform many tasks quite efficiently (Reilly & Cooper, 1990). The information-processing characteristic makes ANNs a powerful technique for prediction problems, since they can learn from examples and then generalise this knowledge to new circumstances.

Many different ANN models have been proposed. One of the most influential models is the multi-layer perceptron model (MLP). This model has the inherent capability of arbitrary input-output mapping and is therefore an appropriate application for forecasting problems. The MLP is typically composed of several layers of nodes, whereas the lowest layer is an input layer where external information is received. Conversely, the highest layer is called the output layer, where the problem solution is obtained. The input layer and the output layer are separated by hidden layers. Figure 6 illustrates a fully connected MLP with a hidden layer and acyclic arcs to connect the layers.

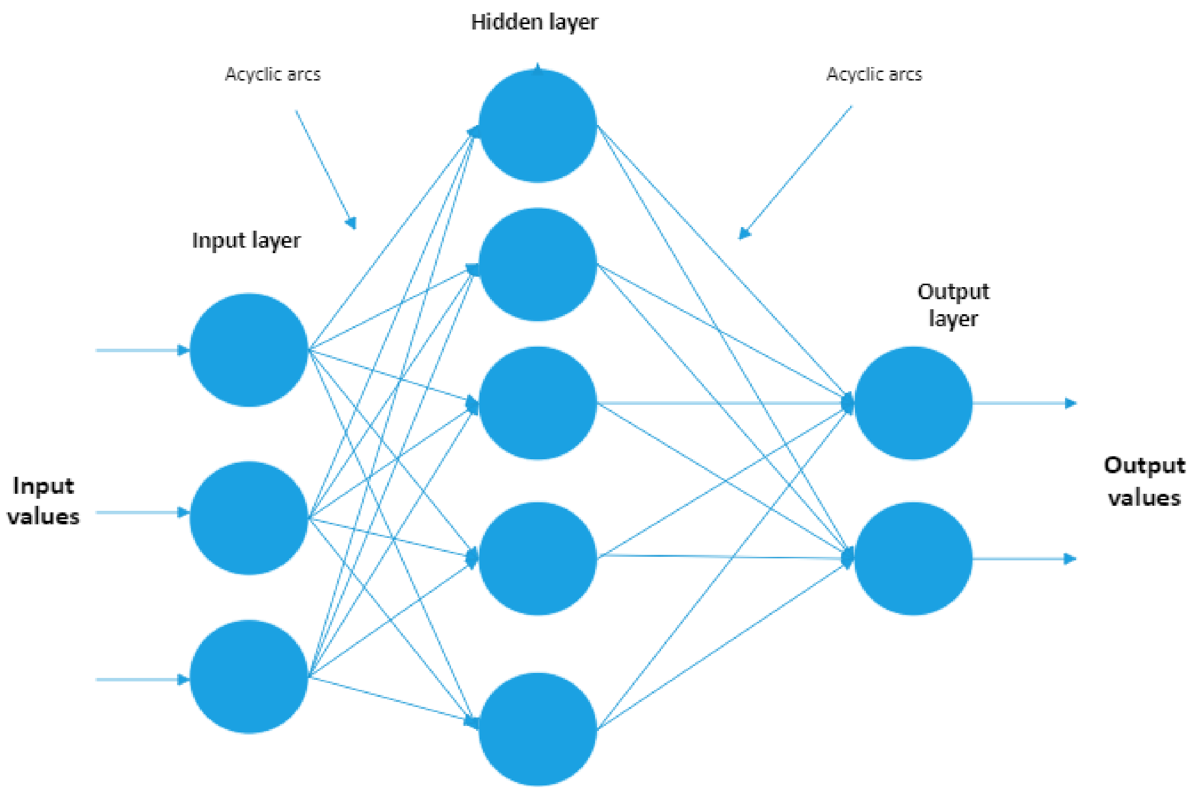

Figure 6 - Typical feed forward neural network (MLP)

For the prediction of the project duration within the utility construction industry, the inputs of the ANN are the independent variables from the dataset. The functional relationship is predicted by the ANN written in equation (2.3), where $x_1, x_2, x_3, \dots, x_n$ represent the n independent variables, and y represents the dependent variable of the prediction.

The artificial neural networks are functionally equivalent to a nonlinear regression model. In addition, considering a project duration forecasting problem, the input data are typically the past observations, and the output is the predicted value. This can be written as the mapping function visualised in equation (2.4), where y_t represents the observation at time t.

$$y = f(x_1, x_2, x_3 \dots, x_n) \qquad \textit{(2.3)}$$

$$y_{t+1} = f(y_1, y_2, y_3 \dots, y_{t-n}) \qquad \textit{(2.4)}$$

Firstly, the total dataset is available and divided into a test dataset, training dataset, and a validation dataset. The test dataset consists of the holdout sample that is used for the test of the model, which focusses on the performance. The training dataset consists of the sample dataset for testing the performance of the model, and the validation set refers to the dataset which is not used for training or testing the model. The artificial neural networks can be used to perform any desired task. However, the ANNs must first be trained before they can perform. This training consists of completing the arc weights of the network, and the knowledge of the function of the network stores are learned by a network in the arcs and nodes in the form of arc weights.

When the training of the network is finished, the model is tested with the testing data to confirm the actual predictive power of the network. However, this only provides a highly rough estimate of how accurate the model will be when presented with new data. Overfitting introduces a major challenge in terms of training the network. This refers to results of weights and biases that generate outputs that perfectly match the training data, which results from the training algorithm running too long. The overfitting will provide a precise match of what is necessary to present the proper results.

Therefore, this research attempts to improve the prediction of project duration at the contractor based on historical data. One of the motivations of the research is the desire of the case company itself to acquire insights regarding their current performance in terms of estimating the project duration. This master's thesis is the first research combined with the contractor's realisation division. The goal is not solely to research the problems at hand but also to provide valuable background research concerning the basics related to the prediction of construction projects.

3.2.2. Background of the regression model

This section elucidates the regression technique. For this research, the regression technique is used as an initial starting point for predicting project durations within the utility construction industry. The regression technique is an independent technique used for predicting project duration and also aims to indicate the relationship between the dependent variable and the independent variable. In this research, the dependent variables are the parameters used for the prediction of project duration in the utility construction industry. The independent variables are the predicted project durations of the projects in the utility construction industry.

To analyse the relationship between the dependent variable and the independent variable, the correlation coefficient is used in the regression analysis. The correlation indicates the relationship between the variables and measures the linear relationship between the two variables. Correlation coefficients invariably have a value between -1 and 1.

3.3. Method for evaluation

The final result of this research is important in relation to the question concerning whether the artificial neural networks can improve the accuracy of utility project duration. This question is evaluated with the help of several performance measurements. The first measurement is the MAE, which consists of the predicted value and the actual value for all observations. Equation 3.1 provides the MAE, which calculates the average absolute deviation. This begins with the MAE, where $\hat{\gamma}_j$ is the predicted value (project duration), and γ_j is the actual value (actual project duration). This determines the difference between the predicted value and the actual value for all n (number) of observations.

$$Mean\ Absolute\ Error = \frac{1}{n}\sum_{j=1}^{n}|\hat{y}_j - y_j| \qquad \text{(eq. 3.1)}$$

The second measurement is the mean percentage error (MPE), which consists of the actual project duration and the predict project duration. The MPE examines the percentage of prediction improvement in a project's duration as a function of the number of orders, n. This measure consists three different metrics: 1 year, 2 years, and 3 years. This approach yields a clear overview of the models, and all of the project durations are measured. The formula of the mean absolute percentage error (MAPE) is provided in 3.2.

$$Mean\ Absolute\ Percentage\ Error = \frac{100\%}{n} \cdot \sum_{j=1}^{n}\left|\frac{y_j - \hat{y}_j}{y_j}\right| \qquad (3.2)$$

The third measurement encompasses the earliness percentage of projects that are finished too early. In the earliness formula, y_j^e represents a binary variable that is 1 if the actual project duration is smaller than the predicted project duration minus the project duration w, divided by 2, and otherwise will be divided by zero. The formula of the earliness percentage is indicated in 3.3.

$$Earliness = \frac{100\%}{n} \cdot \sum_{j=1}^{n} y_j^E$$

$$y_j^E = \begin{cases} 1, & if\ \ y_j < \hat{y}_j - \frac{w}{2} \\ 0, & otherwise \end{cases} \qquad (3.3)$$

The last measurement is the lateness percentage of projects that are delayed. In the lateness formula, y_j^L represents a binary variable that equals 1 if the actual project duration is greater than the predicted project duration. The project duration w is then divided by two and otherwise is divided by zero. The formula of the lateness percentage is presented in 3.4.

$$Lateness = \frac{100\%}{n} \cdot \sum_{j=1}^{n} y_j^L$$

$$y_j^L = \begin{cases} 1, & if\ y_j > \hat{y}_j + \frac{w}{2} \\ 0, & otherwise \end{cases} \quad (3.4)$$

3.4. Data evaluation

This chapter aims to develop an understanding of the data used for analysis in this research. In addition to the characteristics of the data, the background is also described to provide a clear overview of the dataset used in this research.

3.4.1. Data description

The data consists of many columns with information about the projects, including the predicted project duration. To assess whether the project duration is finished in the provided project duration, both the predicted project duration and the actual project duration should be specified in the project. In addition, this section examines the different relevant predictor parameters and how they are presented in the current situation. By using merely the relevant parameters, the final performance of the final model will be higher than when all relevant data is included.

In theory, predicting techniques select the important variables on their own. However, this is tedious and unnecessary at this moment. The human selection of relevant parameters in combination with suggestions from literature should produce a simpler initial starting model. This manner of implication improves the performance of the machine learning algorithm used to derive the final model in neural networks with the tool Matlab. The difference between these methods is that human selection is more based on feeling, and the suggestions from the literature are more based on scientific assumptions.

3.4.2. Data collection and preparation

The dataset obtained from the case study company contains all of the utility projects from January 30th, 2003 to January 30th, 2018. Since all of the relevant data is extracted from the utility department of the case study company, the output is a unified dataset that can be fed into excel or MATLAB. If the data is extracted from different departments, then it was impossible to make the data unified. Furthermore, the regression technique is performed in Excel along with most of the data preparation. The modelling of the ANN has been conducted in the programming language of MATLAB. Therefore, some of the specific data preparation activities necessary for the ANN must be performed in MATLAB.

3.4.3. Data quality and cleaning

This section aims to assess the quality of the dataset. It analyses elements such as missing values, inconsistent values, and typos within the dataset. The data report consists of many columns with specific information about the project. To obtain correct and clear results from the analysis, data errors must be removed from the data. This means that coding errors and missing values must be corrected, and duplicates of the same order must be resolved within the dataset. To do so, the following data cleaning procedures have been applied. A selection has been made to delete missing values, since imputing the missing values would result in greater uncertainty in the data and therefore to less reliable results.

An order within the construction schedule has a starting time for the realisation phase and the end time for the realisation phase. Before the dataset can be used for analysis in this research, data cleaning activities should be performed. The dataset consists of irrelevant information, and this must be excluded from the raw dataset. To exclude the irrelevant data, some filters must be applied to prepare the dataset for analysis. The filters in Table 5 are applied to the raw dataset. In addition, the table describes the percentage of projects that is excluded when applying the specific filter to the raw dataset.

Filter #	Description	Percentage Decline
1	*Construction projects that are not executed in the Netherlands*	**-3.7%**
2	*Renovation projects*	**-84.3%**
3	*Projects that are executed in partnership with contractors*	**-12.4 %**

Table 5 - filters that are applied

4. Case study

The previous chapter presented the techniques selected. This section presents the evaluation of the method on real generated data using a case study at a contractor company based on the demonstration phase of the DSRP. Firstly, the elements of the CRISP-DM framework are used as a guide for performing the case study. It then describes the initial performance at the contractor company. Finally, it develops and tests two models.

4.1. Construction industry

Understanding the utility construction industry is necessary before the data can be used. This section analyses the problem from the perspective of the contractor and underlying business mechanisms. Firstly, a rough understanding of the business case is necessary to explain the goals. Secondly, the target variable is defined. Finally, the process is described to produce an enhanced understanding of the overall process that impacts the accuracy of the utility project duration.

4.1.1. Utility construction industry relevance

Contractors aim at the highest possible quality level for finishing projects in the construction industry. From a business perspective, the reliability of the project in the provided project duration should be 100%. A business could align its production processes with the predicted project duration of the project. Projects that are delayed exert negative effects on the overall process of the company and result in financial consequences.

However, this is infeasible in practice because there are invariably reasons why a project is delayed, such as poor weather or accidents in project. The following section presents the current state of performance within the contractor.

4.2. Current performance

Before the proposed technique is evaluated in a case study, the current performance at the contractor is measured. The analysis yields the results of the different projects using the available data. To provide representative insights in the current performance, 15 years' worth of data is used. In accordance with the manager of construction planner, a specific time frame for data usage has been selected.

The following projects are considered for analysis: office building projects, hospitals or care centre projects and unique projects. The performance is measured based on the different evaluation methods listed in chapter 3.2. The mean absolute error for the planned utility construction projects of contractors are calculated. Figure 7 illustrates the deviation in days of the planned utility construction projects. The average error of the planned utility construction projects by contractors is between 1 day and 23 days.

This indicates that most of the projects are finished within the project duration or too late. From a business perspective, overly late completion of projects is a more severe problem than overly early completion of projects.

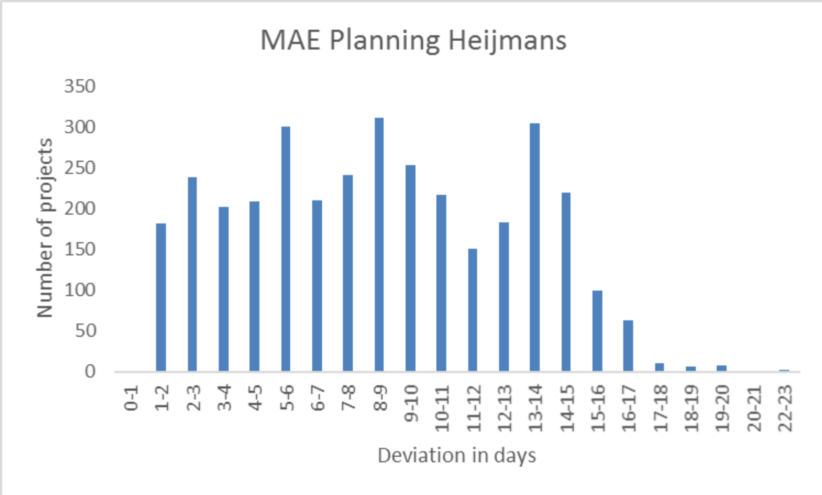

Figure 7 - Mean absolute error of planning Contractor

4.3. Modelling

To compare the techniques used in this research and assess the varying performances, all of the results have been obtained based on the data of the contractor. The models are therefore constructed from data from the period specified previously. In addition, the techniques are validated using 20% of the data relative to the data for building the model. In this research, two techniques for building the prediction model have been selected, namely artificial neural network and regression. These techniques have different variables that have been selected for modelling.

4.3.1. Regression model

The regression technique is based on the variables used as input. For the projects, six variables have been used for predicting the project duration. These parameters have been selected based on the result of the literature review. In addition, the error of the predictions are checked using a normality distribution, as indicated in Table 6. The table indicates the SPSS program terms: degrees of freedom, area sum, average squares, and significance. The validation of the technique results in a slightly higher performance, which is rather unusual. However, this may be the case because the data for validating the regression model is based on a different time period. Therefore, the sample size was sufficient for regression analysis.

ANOVA-tabel					
	Degrees of freedom	*Area sum*	*Average squares*	*F*	*Significance F*
Regression	5	7391526.331	1478305.266	166984.1	0
Faults	3407	30162.0779	8.852972674		
Total	3412	7421688.409			

Table 6 - Results Anova-table

Table 7 shows the model summary. R is the square root of R-Squared and is the correlation between the observed and predicted values of dependent variable. R-Square gives the proportion of variance in the dependent variable (project risks, season, estimate project duration & floor surface) which can be explained by the independent variables (actual project duration). This is an overall measure of the strength of association and does not reflect the extent to which any particular independent variable is associated with the dependent variable. The adjusted R-square is an adjustment of the R-squared that penalizes the addition of extraneous predictors to the model. Standard error of the estimate is also referred to as the root mean squared error. It is the standard deviation of the error term and the square root of the Mean Square for the residuals in the ANOVA table (see table 6).

Model summary	
R	0,997965909
R-square	0,995935955
Adjusted r square	0,99592999
Standard error	2,975394541

Table 7 – Model summary

The mean absolute error for the utility construction project of the planned utility construction projects with regression has been calculated. Figure 8 illustrates the deviation in days of the planned utility construction projects with regression. The average error of the utility construction projects planned by regression are between -8 and 8. This means that the projects are finished between 8 days earlier and 8 days later.

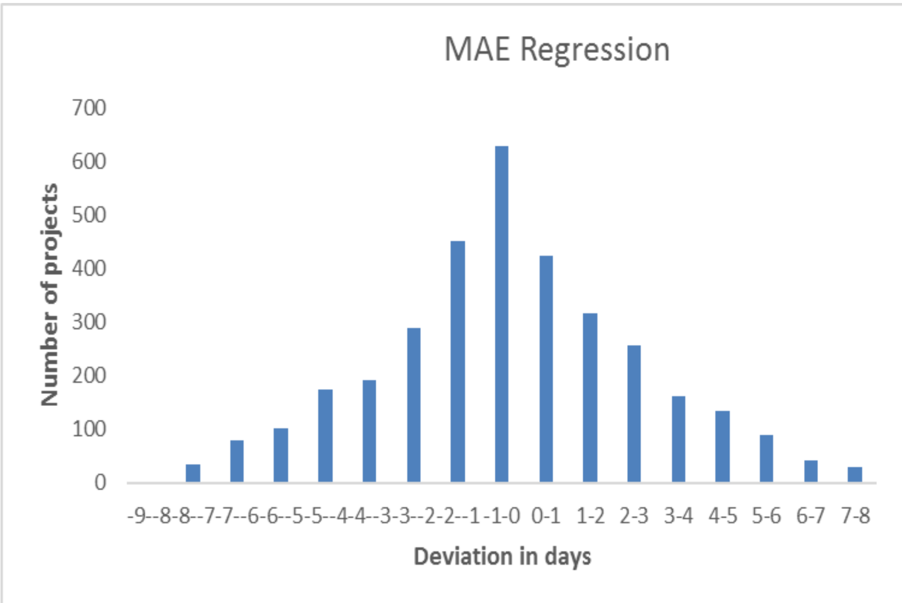

Figure 8 - Mean Absolute Error of the regression

The six variables selected for developing the model have been tested based on their correlation coefficient. As mentioned in Chapter 3, the regression analysis produces important insights. A high positive value indicates a positive linear relationship. In addition, a high negative value indicates a negative linear relationship. This means that a correlation coefficient of zero indicates no linear relationship between two variables. In this research, the same cut-off value is 0.15.

As table 8 demonstrates, the coefficient values mean that there is almost no relationship between the parameters. The variables 'estimated project duration' and 'season start project' indicate a weak positive relationship of 0.025. Furthermore, the variables 'estimated project duration' and 'surface m2' indicate a weak positive relationship of 0.021.

The variables 'surface m2' and 'project risks' demonstrate a weak negative relationship, as table 8 demonstrates. This weak negative relationship is -0.036. The variables 'type utility buildings' and 'surface m2' indicate also a weak negative relationship of -0.012.

	Type utility buildings	**Estimated project duration**	**Season start project**	**Surface m2**	**Project risks**
Type utility buildings	1.000	-	-	-	-
Estimate Project duration	0.016	1.000	-	-	-
Season start project	0.017	0.025	1.000	-	-
Surface m2	-0.012	0.021	0.000	1.000	-
Project risks	0.002	0.003	0.008	-0.036	1.000

Table 8 – Correlation regression

There is no significant correlation between the input parameters (independent variables) and the output parameter (dependent variable). There are no significant correlations among the dependent variables. Merely one correlation indicated a moderate relationship ($0.4 < r < 0.6$). No correlation among the dependent and independent variables means that a promising regression model is unlikely.

4.3.2. Artificial neural networks

The ANN model has three input variables. The input variables consist of the number of input nodes, the number of hidden layers, and the number of nodes in each layer. The number of input nodes is determined using the training module of Levensberg-Marquardt using feature selection with the MAE as performance measurement. The Levenberg–Marquardt algorithm, which was independently developed by Kenneth Levenberg and Donald Marquardt, provides a numerical solution to the problem of minimising a nonlinear function. It is fast and has stable convergence. In the artificial neural-networks field, this algorithm is suitable for training small- and medium-sized problems. There are more training modules, such as the error back propagation (EBP). This training module dispersed the dark clouds on the field of artificial neural networks and could be regarded as one of the most significant breakthroughs for training neural networks. Many improvements have been applied to EBP, but these improvements are relatively minor.

Unfortunately, not all factors that were identified can be used because of several reasons, as mentioned in chapter 3. Therefore, the neural network will be used to predict the project duration based on the historic project duration of projects and by the following six parameters:

1. Type of utility buildings;
2. Estimated project duration;
3. Season start project;
4. Surface m2;
5. Project risks.

Using the project durations of the available output parameters to train the NN yields a larger dataset. To produce a representative output of the performance in this research, the model was run 100 times using the training data. The process in Matlab is presented in a flowchart, as illustrated in figure 9. After the model has calculated the average performance for each input node, the 'best' variable is selected based on 'feature selection' in an excel-file.

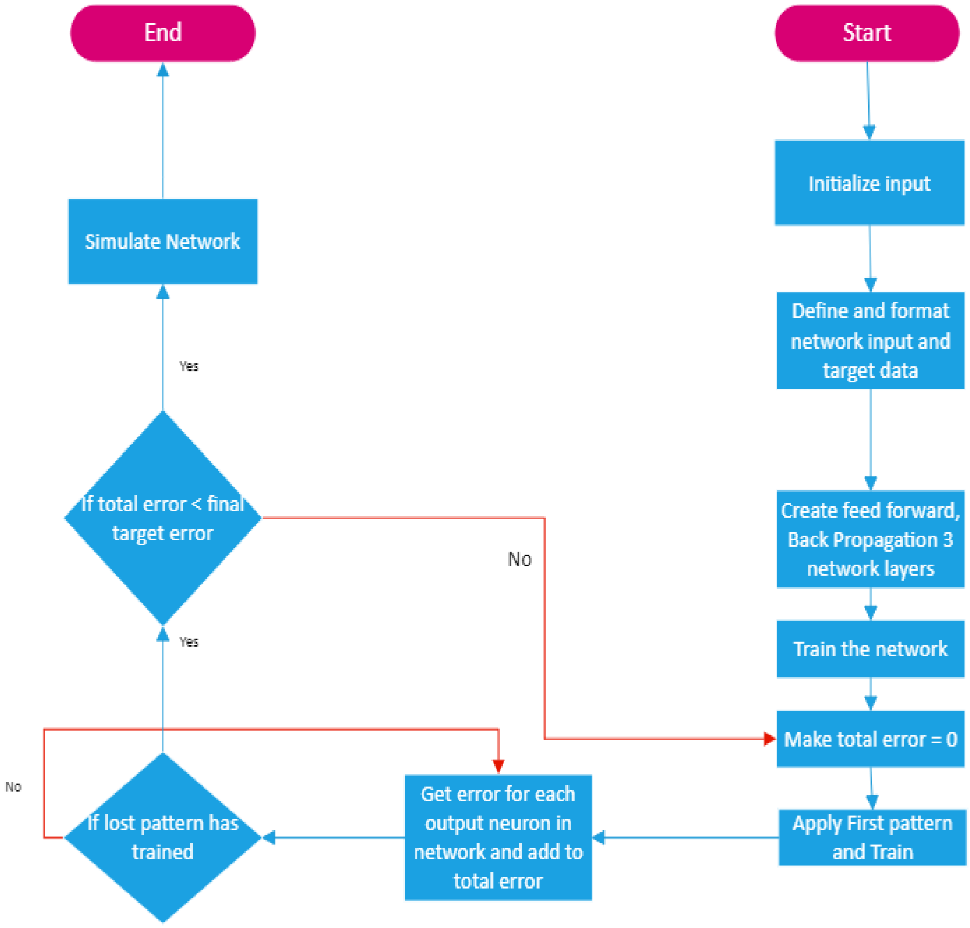

Figure 9 - Flow chart of Matlab

Thereafter, the model runs again in MATLAB with the specific parameters, and the model is optimised by testing different numbers of neurons in each hidden layer in the model. Finally, to produce a representative output, the model is run 10 times in MATLAB, as illustrated in Figure 10.

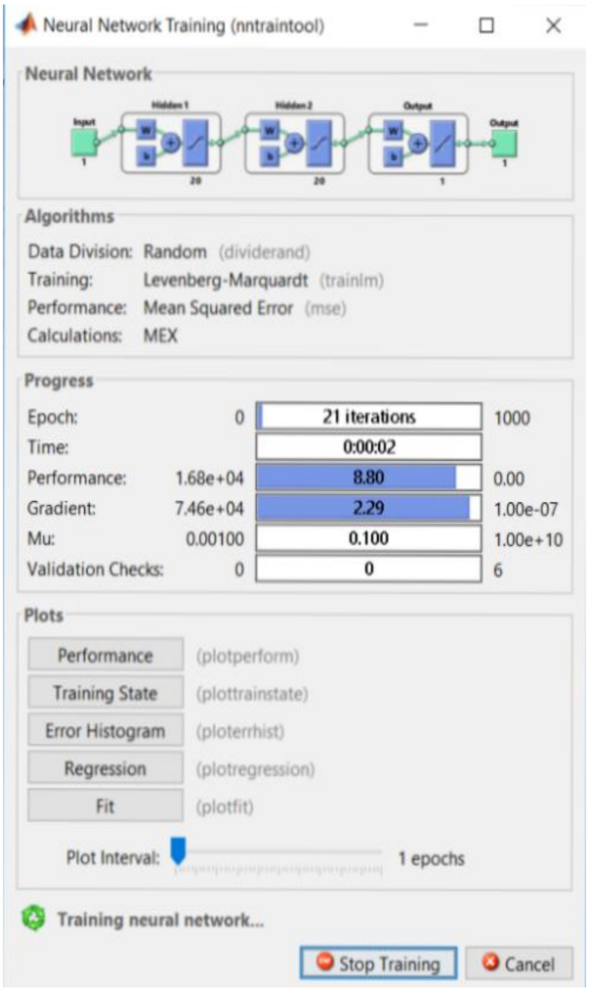

Figure 10 - Run in Matlab

Table 9 represents the performance of the network using different numbers of input nodes and provides the optimal solution. It can be concluded that the optimal number of input nodes is three, namely variables 2, 3, and 4. This results in a model with the parameters ‘estimated project duration’, ‘season start project’, and ‘surface m2’.

Mean absolute error	**2.4**	**2.31**	**2.3085**	**2.5**	**2.8**
Variable	3	3,2	3,2,4	3,2,4,5	3,2,4,5,6

Table 9 – Optimal solution performance of the model

Figure 11 illustrates the performance of the network using different numbers of input nodes in a graph. The figure indicates the MAE of the combination of the optimal variable according to NN with the other variables.

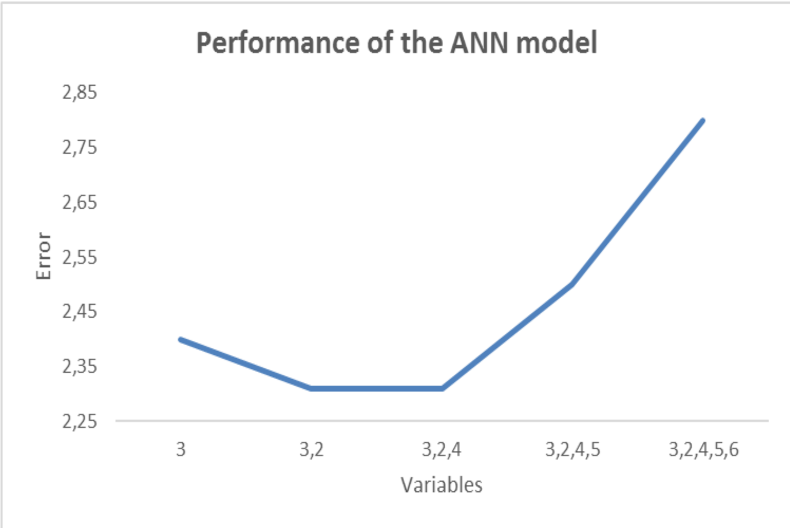

Figure 11 - Performance different input parameters of ANN model

To optimise the performance of the network, the optimal value of neurons in the hidden layers should be determined using parameter tuning. The previous chapter explained the selection of two hidden layers. The optimal number of nodes is determined by testing the performance of the model with different combinations of hidden nodes using the testing data and the optimal number of nodes are variables 2, 3, and 4 . As Table 10 demonstrates, the optimal number of neurons in hidden layers one and two is (20;20)

		Hidden layer 2 size		
Hidden layer 1 size		10	15	20
	10	2.300012	2.322887	2.398419
	15	2.339236	2.357499	2.370751
	20	2.386673	2.356359	2.298731

Table 10 – The performance of the model with different number of nodes in the hidden layer's

The MAE for the utility construction project of the planned utility construction projects with ANN has been calculated. Figure 12 illustrates the deviation in days of the planned utility construction projects with ANN. The average error of the utility construction projects planned by regression is between -9 and 9. This means that the projects are finished between 9 days earlier and 9 days later.

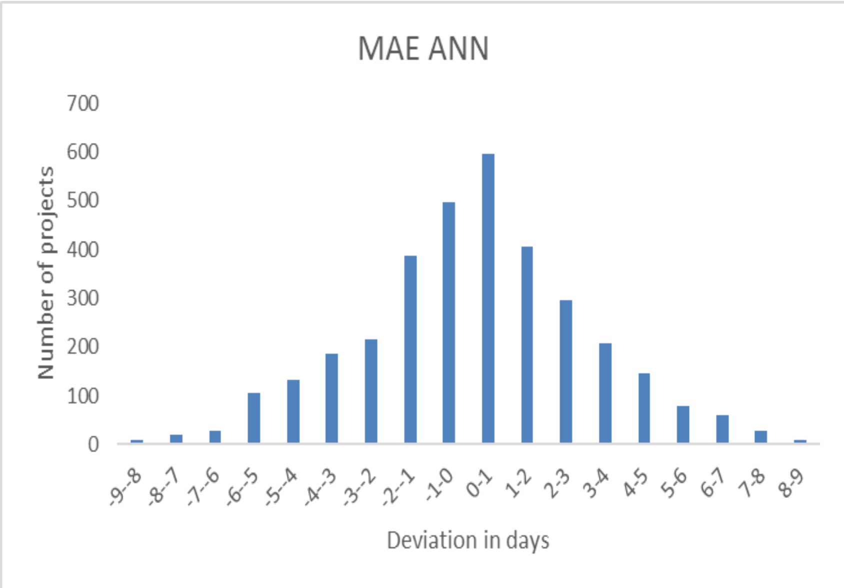

Figure 12 - Mean absolute error of prediction with ANN

4.4. Evaluation and discussion

This section evaluates the performance of the methodology implemented. The previous chapter discussed all of the modelling choices. Before regression and ANN is evaluated in a case study, the current performance at the contractor is measured. The analysis indicates the results of the different projects using the available data. To provide representative insights into the current performance, 15 years of data is used. The following projects are considered for analysis: office building projects, hospital and care centre projects, renovation projects, building transformation projects, and unique projects. The MAE for the utility construction project of the planned utility construction projects of the contractor is calculated. The average error of the planned utility construction projects by the contractor is between 1 day and 23 days. This suggests that most of the projects are finished within the project duration or later.

As section 4.2 demonstrate, the MAE for the planned utility construction projects by the contractor is between 0 and 23. This means that the completion of the planned projects ranges between on-time and 23 days late. Table 11 indicates that 24% of the projects are finished within a deviation of 5 days.

Deviation	Number of projects MAE by contractor
0-1	1
1-2	182
2-3	238
3-4	202
4-5	209
5-6	300
6-7	210
7-8	241
8-9	311
9-10	254
10-11	217
11-12	151
12-13	183
13-14	304
14-15	220
15-16	99
16-17	63
17-18	10
18-19	7
19-20	8
20-21	1
22-23	2
Total (sub)projects	**3,413**

Table 11 – Deviation of projects in the current situation (planned by Contractor)

As section 4.3.1 demonstrates, the MAE for the planned utility construction projects with regression yields a result between -8 and 8. This means that the planned projects are finished between 8 days earlier and 8 days late. Table 12 suggests that 89% of the projects are completed within a deviation of 5 days.

Deviation	Number of projects MAE regression
-9--8	1
-8--7	35
-7--6	80
-6--5	103
-5--4	174
-4--3	191
-3--2	289
-2--1	452
-1-0	629
0-1	424
1-2	318
2-3	258
3-4	163
4-5	134
5-6	89
6-7	43
7-8	30
Total (sub)projects	**3,413**

Table 12 – Deviation of projects with regression

As section 4.3.2 demonstrates, the MAE for the utility construction project of the planned utility construction projects with ANN yields a result between -9 and 9. This means that the planned projects are finished between 9 days earlier and 9 days late. Table 13 indicates that 90.1% of the projects are finished within a deviation of 5 days.

Deviation	Number of projects MAE ANN
-9--8	10
-8--7	20
-7--6	28
-6--5	106
-5--4	133
-4--3	187
-3--2	216
-2--1	388
-1-0	498
0-1	595
1-2	405
2-3	297
3-4	207
4-5	147
5-6	78
6-7	60
7-8	27
8-9	9
Total (sub)projects	**3,413**

Table 13 – Deviation of projects with ANN

The current MAE of the utility construction projects of the contractor is 8.4 days. The MAE of the predicted project duration of the utility construction projects with regression is 9.2 days. The MAE of the predicted project duration with regression increased on average by 9.5% relative to the MAE of the current project duration of utility construction projects. Based on the dataset of the utility construction projects used in this research, the current MAE of the contractor is better than the MAE corresponding with the regression technique.

Furthermore, the MAE of the utility construction projects planned with ANN is 8.52 days. The MAE of the predicted project duration with ANN increased by 1.4% on average relative to the MAE of the current project duration of the utility construction projects. Based on the dataset of the utility construction projects used in this research, the current MAE of the contractor is better than the MAE corresponding with the ANN technique.

Overall, based on the MAE, the ANN technique and the regression technique do not outperform the current situation within the contractor. On the other hand, the MAE for the planned utility construction projects by the contractor yields a result between 0 and 23, which means that the completion of the planned projects ranges between on-time and 23 day late.

The MAE for the planned utility construction projects with regression yields a result between -8 and 8, which means that the planned projects are finished between 8 days earlier and 8 days late. The MAE for the the planned utility construction projects with ANN produces a result between -9 and 9, which means that the planned projects are finished between 9 days earlier and 9 days late. Based on the deviation days and the MAE, the overall conclusion is that artificial neural networks correspond with the best average. This was expected since neural networks have a major application in forecasting. ANNs can learn from a large dataset and recognise patterns easily. The average performance difference between the two implemented techniques is 1.42% for the prediction of utility construction projects in favour of the ANN technique.

Another point of discussion is the quality of the construction project data, as opposed to the quantity. The NN method requires variance in the dataset and diversity. Regarding the problem analysis of Chapter 1, which stated why this master’s thesis was pursued, the project duration can be predicted using a neural network. Neural networks were able to account for all factors identified and for which data was available. However, a proper estimation model could not be developed since the data did not allow the neural network to model properly.

The final conclusion holds that it is difficult to indicate whether project duration prediction can be applicable at the contractor, since the described problems of collecting data of all indicated factors is challenging.

5. Conclusion

This section concludes this report. It first answers the research questions presented in chapter 1 and then discusses the practical implications and recommendations for the case study company. Finally, it addresses the limitations of this study and future research directions.

5.1. Research question

This research aims to improve the accuracy of predictions of project duration for utility construction projects. The main research question was therefore the following:

'How can accurate and reliable historic project durations be used to predict a more efficient construction project duration?'

The first sub-question ('Which methods are currently available for prediction the project duration within the utility construction industry?') was intended to yield an inventory of the different methods that are currently available for predicting the project duration of utility construction projects. This was addressed by conducting a literature study to acquire insights into the different techniques available for predicting utility construction projects. The literature review indicated that the technique that is mostly used for the same projects as utility construction projects is the multi-linear regression. Artificial neural networks are the second most common technique used for the same projects. These techniques form the guidelines for predicting the accuracy of project durations in the utility construction industry.

Furthermore, in the ANN model, the parameters are selected based on feature selection. In the ANN model, the optimal combination of parameters is selected; with these parameters, the final model is developed in Matlab. The performance can be improved using different layers or nodes in the ANN model. This answers the third sub-research question, which was related to making the predicted project duration of utility construction projects more accurate.

The second sub-question ('How can the prediction of project duration in the utility construction industry be made more accurate?') was intended to provide insights into making the prediction of utility construction projects more accurate. To provide fruitful insights, the selected techniques and the selected parameters in the literature review are discussed. The multi-linear regression model tests the relationship between the dependent variable and the independent variables. By adding or deleting parameters from the model, the model could improve the project duration of utility projects. The parameters without data at the case study company and parameters that are not significant are therefore also deleted from the final model. The third sub-question ('What parameters could influence the prediction of project durations?') aims to identify the factors that influence the project durations of utility construction projects.

To assess which parameters should be included in the developed models, the factors that influence actual project durations must be discussed. Using input from the literature review and the experiences of the contractor, the first sub-research question can be answered. The literature review provides a clear overview of which factors are most frequently mentioned in the articles. The three primary factors include late delivery of materials and equipment, weather conditions, and project risks. The factors mentioned in the literature review were converted into parameters to enable the use of the parameters in the created model.

The last sub-research question ('How does the implemented method perform in a case study?') is associated with the case study. To evaluate the different techniques, a dataset was generated to study the performance. First, the initial situation at the case study company was analysed using an actual project duration and a predicted project duration, since this will be used as benchmark. This was accomplished by using the following performance measurements: actual project duration, estimated project duration, correlation, MAE, and MAPE.

In conclusion, based on the MAE, the ANN technique and the regression technique do not outperform the current situation within the contractor. The current MAE of the utility construction projects of the contractor is 8.4 days. The MAE of the predicted project duration of the utility construction projects with regression is 9.2 days, as illustrated in Figure 13. The MAE of the predicted project duration with regression increased by 9.5% on average relative to the MAE of the current project duration of the utility construction projects. Furthermore, the MAE of the utility construction projects planned with ANN is 8.52 days. The MAE of the predicted project duration with ANN increased by 1.4% on average relative to the MAE of the current project duration of the utility construction projects. Based on the dataset of the utility construction projects used in this research, the current MAE of the contractor is better than the MAE corresponding with the ANN technique and the regression technique.

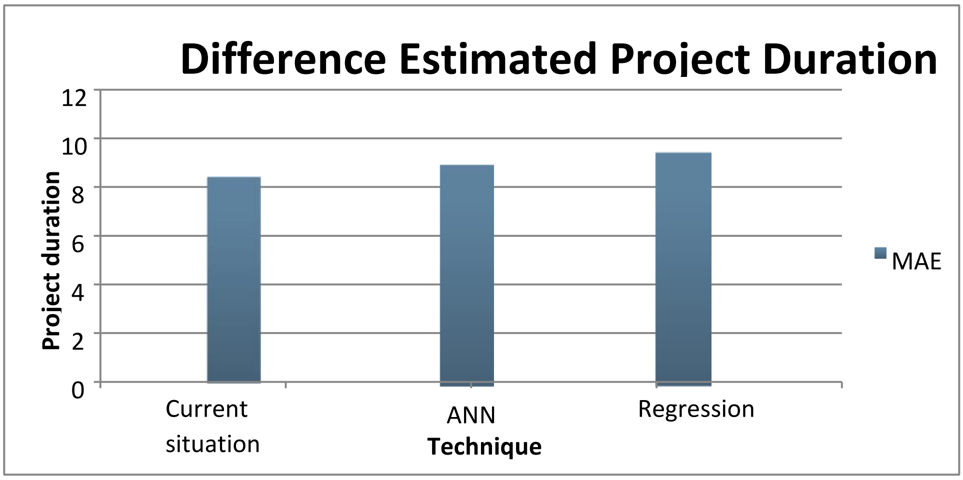

Figure 13 - Mean absolute error of the techniques

By contrast, the MAE for the utility construction project of the planned utility construction projects by the contractor indicates that the completion of the planned projects ranges between on time and 23 days late. The MAE for the utility construction project of the planned utility construction projects with regression demonstrates that the planned projects are finished between 8 days earlier and 8 days late. The MAE for the utility construction project of the planned utility construction projects with ANN indicates that the planned projects are finished between 9 days earlier and 9 days late.

Based on the deviation days and the MAE, the overall conclusion is that artificial neural networks correspond with the best average. This was expected because neural networks have a major application in forecasting. ANNs can learn from a large dataset and recognise patterns easily. The average performance difference between the two implemented techniques is 1.42% for the prediction of utility construction projects in favour of the ANN technique. This is a low improvement in favour of the traditional prediction method.

5.2. Recommendations

The current situation performs significantly better relative to the results of the regression technique and the ANN technique. The neural network model outperforms the regression technique as expected. The conclusion discusses the results of the techniques, which indicate that the artificial neural network technique is the ideal technique to use for predicting project durations. An interesting insight is the fact that when implementing an adjusted project duration of utility projects rather than using the historic project duration of utility projects, the performance approximates the performance of the ANN technique.

The sample size was sufficient for regression analysis. However, more data would be optimal for modelling with neural networks, since more neurons could then be used in the hidden layer, which is ideal for optimising complex prediction problems. Furthermore, the amount of input parameters is sufficient in the sense that it is possible to use neural networks and to perform regression analyses. More input parameters means that more datasets are necessary, which is a problem for NN; this is because when no extra samples are available, this means that the number of neurons in the hidden layer should decrease. This is not ideal, since (based on rules of thumb) using approximately three neurons is recommended. Another recommendation for enhanced findings regarding the input parameters is to implement other input parameters. Unfortunately, not all initially identified factors were possible to use as input for the model used. As mentioned previously, this was associated with several reasons. For most of these problems, it is unlikely that a solution will be discovered in the near future.

Therefore, it is important for the data of the contractor to be structured on the basis of parameters that are necessary to estimate the duration of utility construction projects. If these data are structured, then the ANN technique will be easier to use to predict the duration of utility construction projects.

5.3. Limitations

One of the main limitations in this research is the incomplete nature of the dataset. Therefore, the data of projects within the contractor have not been processed correctly. To conduct this research, it is important to have a useful dataset to test the created models. This research has opted to use the parameters that have data within the contractor for testing the created models. Therefore, for example, it was not possible to apply the dataset of walls or window frames in this research, and the research is consequently not in-depth.

Another limitation is the usage of limited information. Most of the variables that were selected to test in this research emerged from the availability in the initial dataset. Many other variables could influence the utility construction projects; however, these cannot be extracted from the dataset of the contractor. For example, this applies to the specifications of Contractor that the project has selected for prefab construction.

Additionally, an issue with the data is that this is the originally planned data combined with the true project duration. Last-minute changes to the schedule are not accounted for in the dataset. Therefore, the project duration of the utility construction projects is most likely a bit better than presented in this thesis because the adjusted schedule also provides a new project duration which is adjusted for last-minute changes. These last-minute changes in the schedule occur rarely, so the influence on the results is minimal. Finally, the limitation of the new technique is primarily that the technique does not perform effectively for projects that are overly unique. These projects have overly unique specifications, which can mean that no data is available in the big data to use the ANN technique.

5.5. Future research

This study can be regarded as an exploratory study regarding the prediction process with artificial neural networks for contractors in the construction industry. Future research is necessary to develop greater insights into the subject of this study and to fulfil the complete potential of this study.

For future research, several opportunities should be explored. The project duration of utility construction projects currently only relies on knowledge of the planner and some related projects that have been completed in the past. This is insufficient to optimise the project duration within the utility construction industry. A richer dataset combined with an extension of the technique might enhance the already promising results.

Furthermore, new prediction techniques such as the ANN and the regression technique, particularly during the process, can enable smaller project durations for utility projects. Incorporation of these techniques can further enhance the technique and increase the precision of project durations, which can yield a competitive advantage for the contractor.

Bibliography

Abas, Xian, Norfarahayu and Hasmori. (2018). *Investigation on the Factors Influencing Construction Time and Cost Overrun for High-Rise Building Projects In Penang.* Penang, China: Journal of Physics Conference Series 995(1).

Abbasi et al. (2018). *A Decision-Making Framework for Subcontractor Selection in Construction Projects.* Iran.

Adam & Lindahl. (2014). *Implications of Cost Overruns and Time Delays on Major Public Construction Projects.* Sweden: International Symposium on the Advancement of Construction Management and Real Estate.

Al Jurf & Beheiry. (2010). *Qatar residential construction projects.* Qatar: International Conference on Engineering System Management & Applications.

Al-Zarrad, M. A. (2017). *WEATHER DERIVATIVES AS A RISK MANAGEMENT TOOL FOR CONSTRUCTION PROJECTS.* Vancouver, Canada.

Arditi and Pattanakitchamroon. (2006). *Selecting a delay analysis method in resolving construction claims.* Chicago, US: International Journal of Project Management.

B. Samanta & K. Al Balushi. (2001). *Use of time domain features for the neural network based fault diagnosis of a machine tool coolant system.* Muscat, Oman: Proceedings of the Institution of Mechanical Engineers Part I Journal of Systems and Control Engineering.

Balushi, S. &. (2001). *Use of time domain features for the neural network based fault diagnosis of a machine tool coolant system.* Muscat, Oman.

Bhargava, Anastasopoulos, Labi, Sinha and Mannering. (2010). *A Three-Stage Least Squares Analysis of Time and Cost Overruns.* 2010: Journal of Construction Engineering and Management 136(11).

Doree, Holmen, Caerteling. (2003). *Co-operation and competition in the construction industry of the Netherlands.* Twente.

Feiten & Cijfers. (n.d.). *https://www.nam.nl/feiten-en-cijfers.html#iframe=L2VtYmVkL2NvbXBvbmVudC8/aWQ9RGFzaGJvYXJk*. Retrieved from www.nam.nl.

Ghiassi, Saidane & Zimbra. (2005). *A dynamic artificial neural network model for forecasting time series events.* Santa Clara, United States: Operations and Management Information Systems.

Gluszak & Lesniak. (2015). *Construction Delays in Clients Opinion – Multivariate Statistical Analysis.* Poland.

Gorse, Stafford, Miles-Shenton, Johnston, Sutton & Farmer. (2012). *Thermal performance of buildings and the management process.* Leeds.

Gunduz. (2016). *Causes of Construction Delays in Qatar Construction Projects.* Boston, UsA.

Gurgun, Zhang & Touran. (2013). *Schedule contingency analysis for transit projects using a simulation approach.* Istanbul, Turkey: Journal of CIvil Engineering and Management.

K. Movagharnejad & M. Nikzad. (2007). *Modeling of tomato drying using artificial neural network.* Babol, Iran: Faculty of Chemical Engineering, Mazandaran University.

Khosravi, Rahimi, Gheibi & McCartney & Haeri. (2015). *Business Process Intelligence*.

Kikwasi. (2013). *Causes and Effects of Delays and Disruptions in Construction Projects in Tanzania.* Tanzania.

Larsen, Shen, Lindhard & Brunoe. (2015). *Factors Affecting Schedule Delay, Cost Overrun, and Quality Level in Public Construction Projects.* 2015: Journal of management in Engineering.

M. Ghiassi, H. Saidane & D.K. Zimbra. (2005). *A dynamic artificial neural network model for forecasting time series events.* United States: Operations and Managament Information Systems.

Marzouk & El-Rasas. (2014). *Analyzing delay causes in Egyptian construction projects.* Cairo, Egypt: Cairo University.

P.Gonzalez, V.Gonzalez, K.Molenaar & O.Francisco. (2014). *Analysis of Causes of Delay and Time Performance in Construction Projects.* Auckland: Journal of Construction Engineering and Management .

parsonscorp. (2014). *https://parsonscorp.com*. Retrieved from parsonscorp: https://parsonscorp.com/innovation/lean-ready/

Peffers, Tuunanen, Rothenberger & Chatterjee. (2007). *A Design Science Research Methodology for Information Systems Research.* Germany.

R. Jeong & R. Rilett. (2004). *Bus arrival time prediction using artificial neural network model.* Texas, United States: Intelligent Transportation Systems, 2004. Proceedings. The 7th International IEEE Conference on.

Rajguru. (2015). *Effective techniques for finding deley in construction project.* India: Indian Journal of Science and Technology.

Reilly, D., & Cooper, L. (1990). An overview of neural networks: early models to real world systems. *An Introduction to Neural and Electronic Networks* (pp. 227-248). New York: Academic Press.

S. Gardezi, Manarvi & J. Gardezi. (2013). *Time Extension Factors in Construction Industry of Pakistan.* Pakistan.

Said, Benhra & El Hassani. (2013). *Causal Method and TIme Series forecasting model based on Artificial Neural Network.* Morocco.

Samantha, A. M. (2001). *Body mass regulation in response to changes in feeding predictability and overnight energy expenditure.*

Senge, P. (2017). *Thinkipedia*. Retrieved from www.thinkipedia.wikispaces.com: https://thinkipedia.wikispaces.com/What+is+DSRP%3F

Senouci, Ismail & Eldin. (2016). *Time Delay and Cost Overrun in Qatari Public Construction Projects.* Doha, Qatar: University of Houston & Qatar University.

Shehu, Endut and Akintoye. (2017). *Factors contributing to project time and hence cost overrun in the Malaysian construction industry.* Malaysia and UK: Universiti Teknologi MARA and University of Central Lancashire.

Swingler, K. (1996). Chapter 1: Introduction. In K. Swingler, *Applying Neural Networks: a Practical Guide* (pp. 3-21). London: Academic Press Ltd.

Tawil, Khoiry, Arshad, Hamza, Jasri and Badaruzzaman. (2013). *Factors contribute the delay project construction in higher learning education case study.* Malaysia: Research Journal of Applied Sciences, Engineering and Technology.

Thomas NG & Tang. (2010). *Labour-intensive construction sub-contractors: Their critical success factors.* Hong kong.

Tutorialspoint. (2018). *https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_neural_networks.htm*. Retrieved from tutorialspoint: https://www.tutorialspoint.com/artificial_intelligence/artificial_intelligence_neural_networks.htm

Z. Shehu, I. Endut & A. Akintoye. (2014). *Factors contributing to project time and hence cost overrun in the Malaysian construction industry.* Shah Alam, Malaysia: Journal of Financial Management of Property and Construction .

Zairoh et al. (2017). *Influential Factors Affecting Materials Management in Construction Projects.* Malaysia.

Zaitseva & Rogovenko . (2017). *Use of statistical simulation in construction planning.* Russia.

Zhang, G., Eddy Pattuwo, B, & Hu, M. . (1998). *Forecasting with artificial neural networks: the state of the art.* Atlanta, United Status: International journal of forecasting (14), 35-62.

Zidane et al. (2018). *The top 10 universal delay factors in construction projects.* Norway.

Zou, Kiviniemi & W. Jones;. (2017). *A review of risk management through BIM and BIM-related technologies.* Liverpool, UK.

Zou, Zhang & Wang. (2007). *Understanding the key risks in construction projects in China.* China.

Appendices

Appendix 1 – Papers used for appendix 2

Number article	Articles
1	(Abas, Xian, Norfarahayu and Hasmori, 2018) – Investigation on the factors influencing construction time and cost overrun for high-rise building projects
2	(Abbasi et al., 2018) – A Decision-Making Framework for Subcontractor Selection
3	(Adam & Lindahl, 2014) - Implications of Cost Overruns and Time Delays on Major Public Construction Projects
4	(Al Jurf & Beheiry, 2010) - Qatar residential construction projects.
5	(Al-Zarrad, 2017) - WEATHER DERIVATIVES AS A RISK MANAGEMENT TOOL FOR CONSTRUCTION PROJECTS
6	(Arditi and Pattanakitchamroon, 2006) - Selecting a delay analysis method in resolving construction claims
7	(Bhargava, Anastasopoulos, Labi, Sinha and Mannering, 2010) - A Three-Stage Least Squares Analysis of Time and Cost Overruns
8	(Doree, Holmen, Caerteling, 2003) - Co-operation and competition in the construction industry of the Netherlands. Twente.
9	(Gluszak & Lesniak, 2015) - Construction Delays in Clients Opinion
10	(Gorse, Stafford, Miles-Shenton, Johnston, Sutton & Farmer, 2012) - Thermal performance of buildings and the management process.
11	(Gunduz, 2016) - Causes of Construction Delays in Qatar Construction Projects.
12	(Gurgun, Zhang & Touran, 2013) - Schedule contingency analysis for transit projects using a simulation approach
13	(Kikwasi, 2013) - Causes and Effects of Delays and Disruptions in Construction Projects in Tanzania.
14	(Larsen, Shen, Lindhard & Brunoe, 2015) - Factors Affecting Schedule Delay, Cost Overrun, and Quality Level in Public Construction Projects.
15	(Marzouk & El-Rasas, 2014) - Analyzing delay causes in Egyptian construction projects.
16	(P.Gonzalez, V.Gonzalez, K.Molenaar & O.Francisco, 2014) - Analysis of Causes of Delay and Time Performance in Construction Projects.
17	(Rajguru, 2015) - Effective techniques for finding delay in construction project.
18	(S. Gardezi, Manarvi & J. Gardezi, 2013) - Time Extension Factors in Construction Industry of Pakistan.
19	(Senouci, Ismail & Eldin, 2016) - Time Delay and Cost Overrun in Qatari Public Construction Projects.
20	(Shehu, Endut and Akintoye, 2017) - Factors contributing to project time and hence cost overrun in the Malaysian construction industry.
21	(Tawil, Khoiry, Arshad, Hamza, Jasri and Badaruzzaman, 2013) - Factors contribute the delay project construction in higher learning education case study.
22	(Thomas NG & Tang, 2010) - Labour-intensive construction sub-contractors: Their critical success factors.
23	(Zairoh et al., 2017) - Influential Factors Affecting Materials Management in Construction Projects.
24	(Zaitseva & Rogovenko , 2017) - Use of statistical simulation in construction planning.
25	(Zidane et al., 2018) - The top 10 universal delay factors in construction projects.
26	(Zou, Zhang & Wang, 2007) - Understanding the key risks in construction projects in China.

Appendix 2 – Table about influencing factors of utility construction projects

	Articles according appendix 1																									
Factor	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
‘Shortage of materials and equipment in the local market’	O	X	O	O	O	O	X	O	O	O	X	O	X	O	X	O	X	O	O	X	O	O	O	X	O	O
‘Delays in approving permits’	X	O	O	X	O	X	O	O	X	O	X	O	X	O	O	X	O	O	O	X	O	O	O	X	O	X
‘Late delivery of materials and equipment’	X	X	O	X	X	O	X	O	X	X	X	O	X	O	X	X	O	O	X	X	O	X	X	X	X	X
‘Delays in conducting inspections and testing of work’	O	X	O	X	O	X	O	O	O	O	O	O	X	O	X	O	O	O	O	X	O	O	O	X	O	O
‘Delay in monthly payments from owner’	O	O	O	O	O	O	X	O	O	O	O	O	O	O	X	X	O	O	O	X	O	O	O	O	O	O
‘Project risks’	O	X	X	O	X	X	X	O	X	O	O	X	X	X	O	X	O	X	O	O	O	X	X	O	O	O

‘Construction methods’	X	O	O	X	O	X	O	O	X	X	O	O	X	O	X	O	O	O	X	O	O	X	O	O	O	O
‘Incompetent project team’	O	X	O	X	O	X	X	O	X	O	X	O	X	O	O	X	O	O	O	X	O	O	O	X	O	X
‘Design related issues’	O	O	O	X	O	O	O	O	X	X	X	O	X	O	X	X	O	O	O	X	O	O	O	O	O	O
‘Lack of team alignment and conflicts management’	O	X	O	X	X	O	X	O	O	O	X	O	X	O	X	X	O	X	O	X	O	X	O	X	O	X
‘Changed conditions of the project’	O	X	O	X	O	O	O	O	O	X	O	O	X	O	X	X	X	O	O	O	O	O	O	X	O	X
‘Escalation of basic construction material prices in local market’	X	O	O	X	O	O	O	X	X	O	X	O	X	O	O	X	O	O	O	O	O	X	O	O	O	O

‘General labour shortages’	O	X	O	X	O	O	X	O	O	O	O	O	X	O	X	X	O	O	O	X	O	X	O	X	O	X
‘Low labour productivity’	O	O	O	X	O	O	O	O	X	X	X	O	X	O	X	X	O	O	O	X	O	O	O	X	O	O
‘Unrealistic tender cost estimation’	O	X	O	X	O	O	O	O	O	O	O	O	X	X	O	X	O	O	O	X	O	O	O	O	X	O
‘Unrealistic contract duration estimate’	X	O	O	X	O	O	X	O	X	O	X	O	X	O	O	O	O	O	O	X	O	O	O	O	O	X
‘Weather Conditions’	O	X	O	X	X	O	X	O	X	O	X	O	X	O	X	X	O	X	O	X	O	X	O	X	O	O
‘Poor contract administration’	O	X	O	X	O	X	O	O	O	O	O	O	X	O	X	O	O	O	O	O	O	O	O	X	X	O

Legend:

- X: This factor is mentioned in the article
- O: This factor is not mentioned in the article

Appendix 3 – Literature review table about prediction techniques

	Aspects			Prediction technique used to predict the project duration			
Article	Requirement	Implementation in practice	Performance of the technique that are used in the articles	Statistical ANN	Decomposition model with seasonal adjustment	Causal model	Regression
Said et al. (2015)	O	X	O	O	O	X	X
Zhang et al. (1998)	O	O	O	O	X	X	X
Benkachcha et al. (2015)	X	X	O	O	X	O	O
Khosrava et al. (2015)	X	X	O	O	X	O	O
K. R. Al-Balushi et al. (2001)	O	O	O	O	O	X	X
Patel et al. (2016)	O	O	X	O	O	X	X
Remus et al. (1994)	O	O	X	O	X	O	O
Kohzadi et al. (2007)	O	O	O	O	X	O	X

Legend appendix 3	
Implementation column	O: Technique is implemented in practice
	X: Technique is not implemented in practice
Performance column	O: Technique evaluation and performance good
	X: Technique evaluation or performance is not sufficient
Prediction column	O: Tool is used in the article
	X: Tool is not used in the article

Appendix 4 - Script 1 for predicting utility projects

```
%Dividing data into input and output variable
%Dates need to be imported as a datetime type of variable,
not in cell
% termInput(:,1) = Number(:,1);
   termInput(:,2) = Typeproject(:,1);
   termInput(:,3) = EstimateProjectduration(:,1);
   termInput(:,4) = Season(:,1);
   termInput(:,5) = ProjectSize(:,1);
   termInput(:,6) = ProjectRisks(:,1);
   termOutput = ActualProjectDuration(:,1);

   Input2 = termInput(:,:);
    %%%%%%%%%%%%%%%%%%%NEURAL NETWORK
CREATION%%%%%%%%%%%%%%%%
% The amount of times the model needs to be ran
for NumberOfRuns=[1:100];
% To be able to test multiple different inputs in one run
for InputVariables=[2 4:6];
    termInput = Input2(:,[3]);
    %The amount of neurons in layer 1
    for LayerSize1 = [20];
        %The amount of neurons in layer 2
            for LayerSize2 = [20];
     %To see the progress while the program is running
    [NumberOfRuns LayerSize1 LayerSize2 InputVariables]

    x = termInput';
    t = termOutput';

% Create a Pattern Recognition Network
hiddenLayerSize = LayerSize1 + LayerSize2;

%For network with 2 layers
net = fitnet([LayerSize1 LayerSize2]);
%%For retraining
%%net = init(net);

%Amount of layers (+input layer)
net.numLayers = 3;

% Connection between layers
net.layerConnect = [0, 0, 0; 1, 0, 0; 0, 1,0];
%If a layer has a bias
net.biasConnect = [1;1;1];
```

```
% Setup Division of Data for Training, Validation, Testing
net.divideParam.trainRatio = 70/100;
net.divideParam.valRatio = 15/100;
net.divideParam.testRatio = 15/100;

% Train the Network
[net,tr] = train(net,x,t);

% Test the Network
y = net(x);
e = gsubtract(t,y);
tind = vec2ind(t);
yind = vec2ind(y);
percentErrors = sum(tind ~= yind)/numel(tind);

%t2 is the real result y2 is the network output
t2 = t(:,tr.testInd);
y2 = y(:,tr.testInd);
error = t2-y2;
performance = mae(error)


%Store performance with 3 forloops
%PerformanceAvg(LayerSize1,LayerSize2) = performance;
%PerformanceStruct.a{NumberOfRuns} = PerformanceAvg;

%Store performance with 2 forloops
performanceTotal(NumberOfRuns,InputVariables) =
performance;
    end
end
end
end
%Averaging the data for 10 iterations with 2 forloops
performanceTotalAvg = sum(performanceTotal)/100;
min(performanceTotalAvg)
%Convert data into excel date
 y = y';
 %DateStr = datestr((y),' dd-mm-yyyy HH:MM:SS');
%Avg with 3 forloops
% for LayerSize1 = [10 15 20];
 %   for LayerSize2 = [10 15 20];
  %        for NumberOfRuns = [1:10];
   %
SumVector(NumberOfRuns,:)=PerformanceStruct.a{1,NumberOfRu
ns}(LayerSize1,LayerSize2);
    %    end
```

```
% ResultingAvgPerformance(LayerSize1,LayerSize2)
=sum(SumVector)/10;
   %  end
% end
```

Appendix 5 – Script 2 for predicting utility projects

```
%Dividing data into input and output variable
%Dates need to be imported as a datetime type of variable,
not in cell
% termInput(:,1) = Number(:,1);
   termInput(:,2) = Typeproject(:,1);
   termInput(:,3) = EstimateProjectduration(:,1);
   termInput(:,4) = Season(:,1);
   termInput(:,5) = ProjectSize(:,1);
   termInput(:,6) = ProjectRisks(:,1);
   termOutput = ActualProjectDuration(:,1);

   Input2 = termInput(:,:);
    %%%%%%%%%%%%%%%%%%NEURAL NETWORK
CREATION%%%%%%%%%%%%%%%%
% The amount of times the model needs to be ran
for NumberOfRuns=[1:100];
% To be able to test multiple different inputs in one run
for InputVariables=[2 4:6];
    termInput = Input2(:,[3]);
    %The amount of neurons in layer 1
    for LayerSize1 = [20];
        %The amount of neurons in layer 2
            for LayerSize2 = [20];
     %To see the progress while the program is running
    [NumberOfRuns LayerSize1 LayerSize2 InputVariables]

    x = termInput';
    t = termOutput';

% Create a Pattern Recognition Network
hiddenLayerSize = LayerSize1 + LayerSize2;

%For network with 2 layers
net = fitnet([LayerSize1 LayerSize2]);
%%For retraining
%%net = init(net);

%Amount of layers (+input layer)
net.numLayers = 3;

% Connection between layers
net.layerConnect = [0, 0, 0; 1, 0, 0; 0, 1,0];
%If a layer has a bias
net.biasConnect = [1;1;1];
```

```
% Setup Division of Data for Training, Validation, Testing
net.divideParam.trainRatio = 70/100;
net.divideParam.valRatio = 15/100;
net.divideParam.testRatio = 15/100;

% Train the Network
[net,tr] = train(net,x,t);

% Test the Network
y = net(x);
e = gsubtract(t,y);
tind = vec2ind(t);
yind = vec2ind(y);
percentErrors = sum(tind ~= yind)/numel(tind);

%t2 is the real result y2 is the network output
t2 = t(:,tr.testInd);
y2 = y(:,tr.testInd);
error = t2-y2;
performance = mae(error)


%Store performance with 3 forloops
%PerformanceAvg(LayerSize1,LayerSize2) = performance;
%PerformanceStruct.a{NumberOfRuns} = PerformanceAvg;

%Store performance with 2 forloops
performanceTotal(NumberOfRuns,InputVariables) =
performance;
    end
end
end
end
%Averaging the data for 10 iterations with 2 forloops
performanceTotalAvg = sum(performanceTotal)/100;
min(performanceTotalAvg)
%Convert data into excel date
 y = y';
 %DateStr = datestr((y),' dd-mm-yyyy HH:MM:SS');
%Avg with 3 forloops
% for LayerSize1 = [10 15 20];
 %   for LayerSize2 = [10 15 20];
  %        for NumberOfRuns = [1:10];
   %
SumVector(NumberOfRuns,:)=PerformanceStruct.a{1,NumberOfRu
ns}(LayerSize1,LayerSize2);
    %    end
```

```
% ResultingAvgPerformance(LayerSize1,LayerSize2)
=sum(SumVector)/10;
   %  end
% end
```

Appendix 6 - Results of Regression

SAMENVATTING
UITVOER

Gegevens voor de regressie	
Meervoudige correlatiecoëfficiënt R	0,997965909
R-kwadraat	0,995935955
Aangepaste kleinste kwadraat	0,99592999
Standaardfout	2,975394541
Waarnemingen	3413

Variantie-analyse

	Vrijheidsgraden	*Kwadratensom*	*Gemiddelde kwadraten*	*F*	*Significantie F*
Regressie	5	7391526,331	1478305,266	166984,1	0
Storing	3407	30162,0779	8,852972674		
Totaal	3412	7421688,409			

Appendix 7 – Performance of the model with different number of nodes in the hidden layer's

										10					15					20
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	2,300012	0	0	0	0	2,322887	0	0	0	0	2,398419
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	2,339236	0	0	0	0	2,357499	0	0	0	0	2,370751
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	2,386673	0	0	0	0	2,356359	0	0	0	0	2,298731