A Predictive Model for Scaffolding Manhours in Heavy Industrial Construction Projects: An application of machine learning

Kavana Siddappa

A Thesis In

The Department

Of

Building, Civil and Environmental Engineering

Presented in Partial Fulfilment of the Requirements

For the Degree of

Master of Applied Science in Building Engineering at

Concordia University

Montreal, Quebec, Canada

Winter 2019

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared by

By: Kavana Siddappa

Entitled: A Predictive Model for Scaffolding Man-hours in Heavy Industrial Construction Projects: An application of machine learning

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Building Engineering)

Complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

Prof. Joonhee Lee, Chair

Prof. Fuzhan Nasiri (Examiner)

Prof. Mazdak Nik-Bakht (Examiner)

Prof. Sang Hyeok Han, Supervisor

Approved by Dr. F. Haghighat

Chair of Department or Graduate Program Director

March 2019 Dr Amir Asif

Dean of Faculty

Abstract

In cold countries like Canada, modular construction is widely adopted in heavy industrial construction projects due to weather uncertainties. To facilitate the construction processes, the temporary structures, especially scaffolding, are essential since it provides easy access for workers to carry out construction activities at different levels of the height and also ensures the safety of labourers. As indirect costs of projects, the scaffolding is estimated by 15-40% of project costs. Furthermore, according to increase the size of the projects, the scaffolding uses larger amount of resources than estimated ones, which may cause budget overrun and schedule delay. However, due to the lack of systematic and scientific models to estimate the scaffolding productivity, heavy industrial company has difficulty to plan and allocate the resources for scaffold activities before construction. To overcome these challenges, this paper proposes a predictive model to estimate scaffolding productivity based on the historical scaffolding data of a heavy industrial project. The proposed model is developed based on the following steps: (i) identifying the key parameters (e.g. specific trades, work type, different scaffold methods, task times spent using scaffolds, and weights of the scaffolds) that influence the scaffolding manhours and project productivity; and (ii) developing the predictive models for scaffold manhours using machine learning algorithms including multiple linear regression, decision tree regression, random forest regression and artificial neural networks(ANN) . The accuracy of models have been measured with evaluation metrics which are mean absolute error (MAE) and root mean squared error (RMSE) and the R squared value. The findings reveal upto 90% accuracy for ANN models.

Acknowledgment

I would like to thank my supervisor Dr Han, for giving me immense support all through this research process. I would be grateful for the PCL Company and its employees for helping me with providing case studies for my research. The timely guidance from both professor and the PCL Company helped me to complete the research on time. Further, my family and friends are of great support in this process. Their constant support have made me more dedicated towards my work. I would also want to thank Concordia University for giving me this chance to follow my dream and providing accommodation to pursue my graduation degree.

Table of Contents

List of Figures

List of Tables

CHAPTER 1: INTRODUCTION

1.1. Back Ground

Construction Industry is one of the major driving tools of the Canadian economy. Most of the economy for construction is highly dependent on heavy industrial projects such as petroleum, mining, shipbuilding, steel, chemicals, machinery manufacturing, and oil refineries since they involve huge capital and resources. A forecast for the upcoming decade (2019-2028) conducted under the assistance of Government of Canada, says that the mining, quarrying and other oil sand industries would be the major influence for driving the construction economy [1]. In particular, Alberta's oil sands stands as a primary source of growth in Canada's heavy industrial projects. As for the scale of development, based on the oil sands market analysis by Ernst and Young [2], the impact of the oil sands on Canada's economy is forecasted to reach a total of nearly $4.93 trillion for the years 2010 to 2035, and over 90% of the economic impact will be felt in Alberta. Meanwhile, according to the Alberta oil sands supply chain opportunity analysis [3], total capital expenditures on oil sands projects are forecasted to exceed $150 billion over the next 10 years. The range of estimated total expenditures on maintenance, repair, and operation between 2011 and 2022 is from $227 billion to $330 billion [3]. Based on the actual figures and projections, steel fabrication and machinery manufacturing are expected to play a critical role in oil sands projects. In addition, the statistics of Canadian construction claims that the main challenge construction industries would face in the near future would be labor resources. The government anticipates that around 2 million laborers are in a verge to retirement and the demand for labor would go high, since most of the industrial works happens in remote areas where there is less availability of human resources. Also, there is a scarcity of young laborers who are willing to work for heavy industrial projects [4]. These kinds of projects involve heavy equipment such as boilers, pressure vessels, tanks, heat exchangers, and

steel pipes and tubes which are designed for many disciplines of work (structure, civil, chemical, manufacturing, and mechanical). To facilitate the construction of these components, each discipline needs the temporary structures amongst which scaffolding is widely used. Scaffolding provides temporary elevated platforms, thereby allowing laborers to access their work areas and transport materials vertically and horizontally. In this respect, the scaffolding in the industrial projects should be installed, modified and/or dismantled in accordance with the requirements of the disciplines on their demand times in order to prevent project schedule delays. Due to the demand-based scaffolding operation, the construction domain has difficulty to plan scaffolding operation in the early phases of the project. This has led to ad hoc way of using resources such as labor, equipment, and materials. There is a need for effective utilization of resources, especially the manpower because the labor cost is highly expensive and there is a lack of human resources.

1.2. Research Motivation

The day-to-day challenges faced in the construction processes provide the room of opportunity for improving the work productivity by adopting new technologies and/or by implementing various innovative approaches. However, it is believed that the complexities in heavy industrial projects involving larger machine tools, huge facilities, and various work processes are higher than the other construction sectors (e.g., residential, civil and commercial). The reason behind the complex nature of the industrial sector is; it involves higher risk in handling materials and laborers since they are bulk projects. There is a need of proper planning in order to avoid the uncertainties such as overflow of time, cost and quality may arise and affect the overall project's performance throughout its lifecycle. Hence, an effective planning is of utmost importance for heavy industrial projects.

Over the years, planning of the projects, primarily concentrated on permanent structures, as they tend to consume more workforce and resources. The permanent structures such as

buildings, bridges, and tunnels can be aided through certain erected temporary structures by providing access, support, and protection for the facilities which are under construction. Apart from this, temporary structures are also used in various above and below ground facilities in order to facilitate inspection, repair and maintenance works [5]. Some of the examples of temporary structures are earth-retaining structures, tunneling supports, underpinning, diaphragm/slurry walls, construction ramps, runways, and scaffolding. Once the permanent structures are completed, these temporary structures are either incorporated along with them or are broken into pieces and separated [5]. The current industrial practices are highly dependent on the knowledge and experiences of individual engineers for most of the temporary works. There is a lack of planning and estimation for these temporary works prior to the start of the project. The current market has a lot of software designed for temporary facilities (e.g., shoring and scaffolding) which are commercially available. Unfortunately, all these are not globally adopted due to their limitations of not being implemented successfully in any kind of working environment. The main decisions have been taken by human cognition, based on the visualization of building designs or construction sites [6]. Many construction projects use temporary facilities regularly, and the safety, quality, profitability, and duration of these structures are having a larger impact on the overall budget and time of any construction project considered. They sometimes exceed more than a quarter percent of the total cost of the project. Hence, the careful study and planning of these activities is a requirement for current construction industry's scenario. Generally, construction projects require different kinds of temporary works which consume a considerable amount of time and cost of the project, one such work is scaffolding. The recent discoveries in the field of project management proves that the scaffolding works do consume a substantial amount of resources (e.g., labours, equipment and space), which sometimes could create budget overrun, schedule delay and other issues such as safety, quality, and profitability of the projects when they are not planned efficiently in the

planning stages of the construction projects [7]. Scaffolding has been an essential temporary work, which is required in both prefabrication sites and in the construction sites for providing temporary elevation platforms to facilitate the labor works as well as material transferring. In practice, the companies roughly estimates the scaffolding manhours by a considerable percentage (15-40%) of the total manpower of the project, which is a significant portion of the total cost of the project. Since there are no standardized procedures for guiding the scaffolding practitioners in regard to planning the scaffold works, the man power may exceed the range quoted by scaffolding expertise and may cause cost overruns and/or schedule delays. Furthermore, previous researchers and the industrial companies hardly gave attention to scaffolding and developed any scientific and systematic models for estimating scaffolding manhours. .Henceforth, there is a necessity of developing scaffold-planning models for not only efficient execution of the projects but also project productivity improvement in terms of cost and time. In this respect, this research helps to build a predictive model for scaffolding manhours in heavy construction projects based on the application of machine learning algorithms including multiple linear regression, decision tree regression, random forest regression and artificial neural networks (ANN). By the effective utilization of a case study obtained from a collaboration company, the proposed model is developed and validated in the planning stages of construction projects.

1.3. Objectives

The research objective is to know which all factors are impacting the manhours consumption by scaffold activities and how much percent of manhours have been used for scaffolding works for the heavy industrial projects. The traditional practice for manhours required for each scaffolding task is decided based on scaffold expertise which is subjective based on their experiences. The decision may vary from person to person. Deciding the work schedule and allotting the resources for scaffolding activities might vary for different reasons such as delay

of the work due to labor /material unavailability, or the temperature may not be favourable and many.. Some of the subcontractors prefer scheduling the scaffolding activity based on the trade of work (giving priority to a particular trade) whereas some subcontractors use the date of request raised (the early request is considered first) for conducting the work. This might lead to a lot of impromptu decisions. To avoid ad hoc methods of estimating and allotting resources (such as man power and materials), there is a need to plan for scaffolding activities in prior. Hence, predicting the manhours for scaffolding activities during the planning phase of the project in order to avoid unnecessary cost overruns and schedule delays is considered as the main goal of this research. In this way, the supervisor can have an idea of manpower requirement and plan the associated works in a much detailed way. The proposed method would estimate and keep track of the manhours required for scaffolding works using technical methods and will further help during the decision-making process for planning efficiently. A case study of the industrial project from a well-known heavy industrial construction company has been considered for this research. The study is undertaken step by step in the following process.

i. Study of existing scaffolding process and the requirements by referring to historical data of the industrial projects.

ii. Data visualization in order to understand and monitor utilization of scaffold data in relation to manhours consumed for scaffolding tasks.

iii. Development of a predictive model for scaffold manhours using machine learning algorithms based on the previous project data analysis in terms of associated specific trades, scaffold weights/volumes, work type, and other impacting factors.

1.4. Thesis Organisation

Chapter 1 deals with the back ground and motivation of the research .The importance of temporary structures, especially scaffolding activities in the heavy industrial companies, the consequence of having improper planning of scaffolding are discussed.

Chapter 2 deals with the literature review associated with the scaffolding. Types of scaffolds, planning and existing methods of scaffold estimation, automation in construction field, and implementation of machine learning in the construction field are critically reviewed.

Chapter 3 deals with the general method of data analysis process and how it is implemented to current research for building a predictive model using machine learning algorithms. A step by step process of data analysis is explained. Starting from collection of scaffold data, pre-processing by feature selection, removing outliers, data visualisation through graphs and correlation charts are explained in detail. Further, various machine learning algorithm such as Multiple Linear Regression, Decision Tree Regression, Random Forest Regression and Artificial Neural Networks (ANN) are discussed. Also, evaluation metrics such as Root Mean Squared Error (RMSE); Mean Absolute error (MAE), and R squared values are discussed.

Chapter 4 deals with the results of the predictive model built using machine learning algorithms with the help of existing data. The test results of each model is tabulated and compared. The results are represented in terms of accuracy and error values. A brief explanation

Chapter 5 deals with the conclusion of the whole research and scope for future works. It also addresses the limitations for the current methodology

CHAPTER 2 LITERATURE REVIEW

This chapter explains about the basics of scaffolding and different types of scaffolding structures used in the field of construction. In addition, it describes about the recent developments in the field of construction in terms of automation, application of machine learning and various other innovative methods. Further, this section put lights on the existing methods of scaffold planning and estimation and it also explains the approach of this research.

2.1. Introduction to Scaffolding

Occupational safety and health services (OSHA) defines scaffolding as any structure, framework, swinging stage, suspended scaffolding, or boatswain's chair, of a temporary nature which are used to support and/or protect construction workers by providing easy access to construction work areas horizontally and vertically, and also helps in material transfers [8]. Based on the loads to be carried on the scaffolding structure, materials such as bricks, steel, blocks, bamboo are chosen to build scaffolding [9]. There are different components to build scaffolding. It varies for each type of the scaffolding. However, there are some basic components required to construct any type of scaffolding, which includes base plates or castors, mudsills, adjustable screw jacks, vertical braces on both sides of frames, horizontal braces on every third tier of frame, platform materials to deck the working level, guardrails with toe boards, guardrail posts, ladders or stairs for access and intermediate platforms [10]. Choosing the components for building a particular scaffold is upto the manufacturer's choice. But, selecting the right scaffold for the work is a technical concern, since it depends on various factors such as

i. Site conditions- As in concrete floors, exterior, interiors, backfill.
ii. Weight of workers, materials, tools and equipment's that has to be carried.
iii. Anticipated weather conditions.

iv. Height at which the work is carried.

v. Experience of the crew.

vi. Type of work such as for painting, electrical, piping for which the scaffold is built.

vii. Duration of work [10].

It should be noted that there are different kinds of scaffold used by the construction industries based on the requirements mentioned above. In North America and Canada, generally the construction industries prefer certain types of scaffolds. Each of the scaffold are explained in brief.

2.1.1. Tubular Frame Scaffolds

Tubular frame scaffolds are the standard scaffolds in the construction industries. The main advantage is the frames are available in different sizes and configurations, and also they are easy to assemble and can be done manually. They are generally made of steel but in the current practices, aluminium is chosen [10].An example of tubular frame scaffold is represented in Figure 1.

Figure 1 Example of tubular frame Scaffold [11]

2.1.2. Tube and clamp scaffolds

Tube and clamp scaffolds, also known as tube and coupler scaffolding are one of the oldest type of scaffolding which emerged its way into the construction industry back in early 1900s and is still in use because of the fact that they are made of steel, which adds strength to the structure and also they are easy to assemble and dismantle. This type of scaffolding is the most flexible type of built-up scaffolding because it is usable to any type of configuration, as in it can be used in any direction such as vertical, horizontal or diagonal. Due this flexible nature, it is used for extremely complex designs. The only main disadvantage of this tube and clamp scaffolding is, it uses more labor force the high consumption of labour force which extends the cost consumption [9]. However, it is highly recommended to build these kind of scaffolds with the help of a scaffolding expertise to ensure the safety of the workers. Also, after certain height (10m), it is a rule that this scaffold has to be designed by professional engineer [10]. An example of tube and clamp scaffold is represented in Figure 2.

Figure 2 Example of tube and clamp Scaffolding [12]

2.1.3. Systems scaffolds

Systems scaffolding which are also referred as modular system scaffolding are systematically built with the help of horizontally, vertically and diagonally pre-engineered components of fixed incremental lengths. An important feature of all the types of system scaffolding is, the device with which it connects all the horizontal ledgers to the vertical nodes. Most of the features that contributes towards the importance of this scaffold type are safety, speed, efficiency and consistency. Another contributor to its fame is, its simplicity in terms of assembling which ensures that the amount of loose pieces are fewer when compared to other assembling scaffolding systems. This scaffolding is a bit more expensive than the other types because its initial investment cost is high [9].Though the height and width of the scaffold cannot be adjusted as in tube and clamp, systems scaffolds are widely adopted for non-rectangular, dome-shaped and circular structures [10].An example of the system scaffolds is represented in Figure 3.

Figure 3 Example of System Scaffolding [13]

2.1.4. Tower Scaffolds

Tower scaffolds also known as staircase towers are the scaffolds built for the safety of labor working at heights. It’s always designed by an engineer since its failure in erecting can cause serious damages. The interesting thing about tower scaffolding is that, it's not necessary for it to be built from the ground level. Based on the requirement, it can be built from any height

above the ground level along with the secure access. Such type of scaffolds are useful in situations where the customers might want the structures to not meet with the ground and obstruct the path for pedestrians. The key features of this type of scaffold is -its fast to build, certain types of tower scaffolds can be built by a single person and few might not even require any kind of prior experience [9]. Tower scaffolds are widely used in North America and Canada. The entire scaffold can be raised and released to the required height providing the workers a comfortable working platform. The manufacturer's regulations should be always followed to build the tower scaffold [10].An example of tower scaffold is represented in Figure 4.

Figure 4 Example of Tower Scaffolding [14]

2.1.5. Dance floors

Dance floor scaffolds are temporary platforms erected along with the regular type of scaffolding. They are built in a way that multiple workers can work in the same platform. This in turn helps to complete the tasks in a shorter span. They are generally used for ceiling works and lobby areas [15].An example of dance floor type of scaffold is represented in Figure 5.

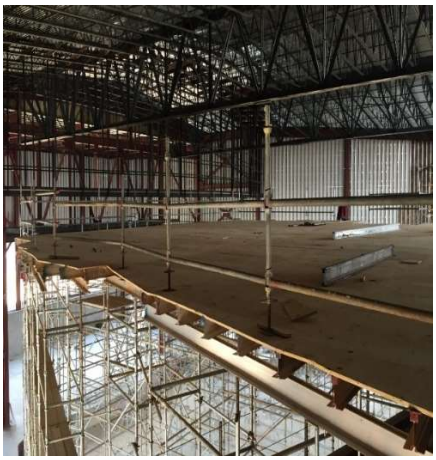


Figure 5 Example of Dance floor Scaffolding [15]

2.1.6. Cantilever Scaffolding

Cantilever scaffolds are special kind of scaffolding which are used when there is difficulty in placing the base of the scaffold on the ground or in the congested pathways. The use of poles and frames are not required in this kind of scaffolding and also they are easy to erect [16].An example of cantilever scaffolding is represented in Figure 6.

Figure 6 Example of Cantilever Scaffolding [16]

2.1.7. Hanger Scaffolding

Hanger scaffoldings are the suspended types of scaffolding which are built using ropes and pulleys. This is one of the scaffolding used by the workers during maintenance activities for moving above and below the ground level [17].A hanger scaffolding example is represented in Figure 7.

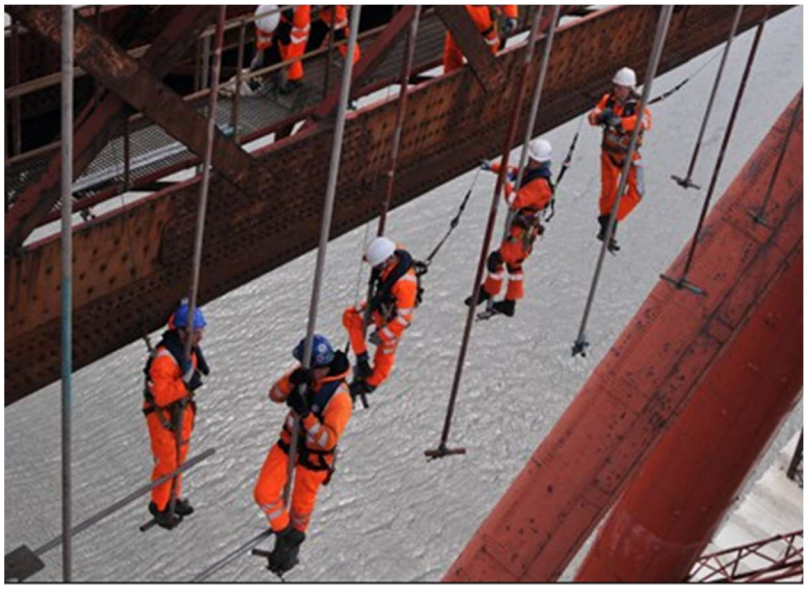


Figure 7 Example of hanger suspended Scaffolding [18]

2.2. Need of Scaffold Planning in Industrial Sector

Being one of the largest sector of the construction domain, the heavy industrial projects lack obligation of having proper scheduling and planning throughout its life cycle. Industrial projects do have their complexities and uncertainties due to the involvement of larger number of man power, labor resources and many other complicated process. In practice, the estimation of the cost for any construction project is generally divided into two types of cost; namely Direct and Indirect costs. The direct costs include labor, materials, supplies, equipment, and any expenses related to the final product. The indirect cost comprises of overheads, profits, and contingence allowances and other temporary costs that does not fit in the Work Break down Structure (WBS) of the project. Surprisingly, these indirect costs have exceeded up to 55% of the total project cost [19]. There are different ways of estimating the project costs depending on engineering experiences and knowledge, and regulations in companies. By close investigations it can be said that the estimator's experience and knowledge is the key deciding factor for estimation of indirect works. Most of the industrial companies claim that the percentage of indirect cost required for scaffolding works earlier used to vary from 15-20% of

the total direct work of the project, but it is extended to 30% and more which is not at all a negligible amount [7]. Since there are no scientific and objective procedures to guide scaffold expertise in regards to planning and estimating of scaffold work, uncertainties such as cost and schedule overruns have become more common. About the schedule delays; the work phase model created by the Construction Owners Association of Alberta (COAA) do suggest to have an effective planning and scheduling for scaffold activities since any delay in the scaffold work would affect the other major works directly [20]. COAA recommends having integral planning for scaffolding in each work packages of any construction project. By collective thoughts put together, it can be said that, scaffolding accounts for considerable percentage of overall cost of the project. However, to overcome the current issues (e.g., cost and schedule overruns) faced by scaffolding activity, due to potentially incorrect human judgments or decisions, there is a need of proper planning and estimation of scaffolding activities in the planning stages of construction projects. It is urgent to develop a scientific model and/or method for scaffolding planning and estimation [7].

2.3. Current Scaffolding Practices

Planning of scaffolding activities is completely subjective and differs from company to company. It is worth mentioning that, most of the companies plan scaffolding as on need basis that is day–to-day or weekly basis [7]. For any particular project, generally the scaffolding superintendent is in charge for allotting the scaffold works. Initially, the trade foremen raises a request for the scaffold, then the superintendent will approve the request based on the availability of the resources (e.g. labor and materials) [7]. An industrial company claims that; in some scenarios, the scaffolding supervisor consider the early request dates, as in the first requested scaffolding task is approved for the construction. However, sometimes there decisions are taken based on the trade of work for which the scaffolding is required immediately. It is completely subjective decision to supply the scaffolds to the requested site

[7]. Figure 8 describes a general method of scaffold workflow in practice. Further, the allocation of manhours to accomplish any scaffold tasks is also decided by various factors such as the scaffolding weights or volumes, the type of scaffold to be built, the trade for which the scaffold is built. It can be summarised that the whole scaffolding activities has been an individualistic approach when it comes to scaffolding planning. However, this subjective decision often leads to excessive use of labor, schedule delays due to non-approval of scaffolding requests, resource shortages and in turn affect the overall project cost. An advanced plan of scaffolding activities would help in the effective use of the resource such as labor force and scaffolding materials. Further, a prior planning also helps in coordination between different project management teams within the project which helps to take right decisions in terms of time and cost [21]. Retrieving the necessary information and automated tracking of on-going construction works has proved to be a better solution; for taking decisions during the planning of work processes and the ability to handle dynamic circumstances [22].

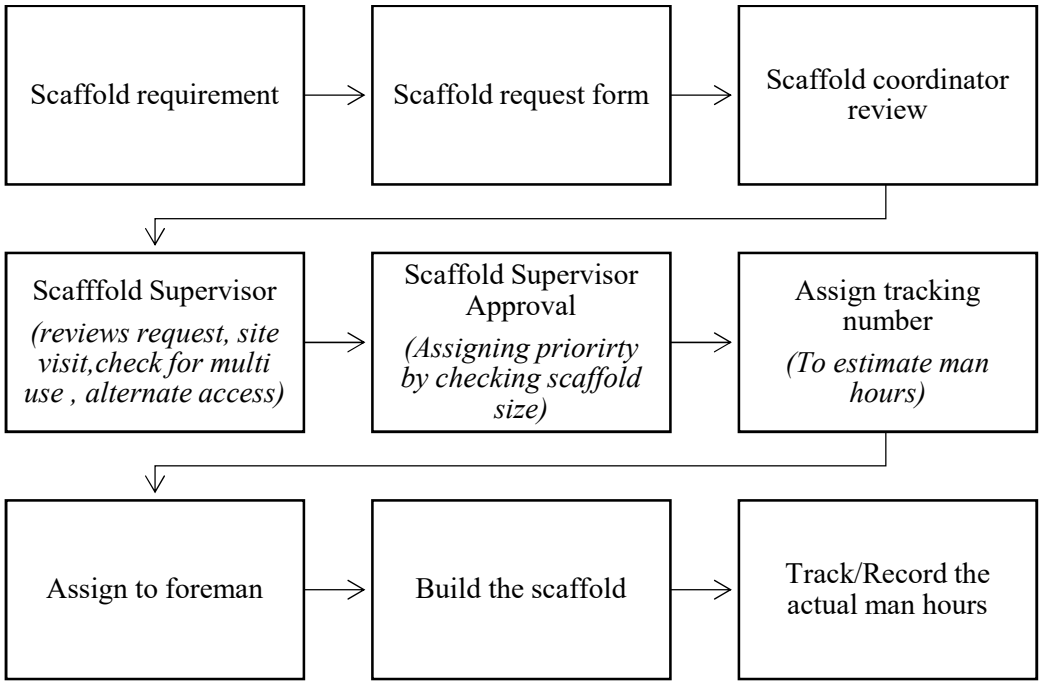


Figure 8 Work flow of a single scaffolding activity

2.4. Machine Learning (ML) Algorithms in Construction

The traditional practices of handling construction projects are by visual inspection and human cognition. However, the recent developments in the field of science and technology has led to many data driven software and other innovations in the construction field [22]. Every aspect of the construction is getting adapted to new software, whether it might be scheduling, estimation, safety analysis or mitigating risks [22]. Through machine learning techniques, existing data can be effectively used in building automated monitoring systems through machine learning [22]. Use of machine learning helps to analyze the data of the construction works, which would further provide effective insights and guides the engineers/project managers during decision-making phases. Time, cost, quality, safety, operations and maintenance, and many other aspects of project management have adopted machine learning (ML) algorithms [23]. There are four different types of algorithms that can be used to project management areas for planning and estimation; they are classification, regression, association, and clustering [24]. Both classification and regression are used as predictive models where classification has categorical outputs and regression has continuous variable has outputs. Association is used for finding the relationships between the variables and clustering is a segmentation process where the similar variables are grouped [24]. In works related to construction, a lot of attempts has been made to adopt machine learning techniques. For instance, in 1997, Siquera [25] implemented machine learning algorithms (neural networks and regression models) for performing cost estimation to low rise buildings. This was one of the initiatives taken towards the emergence of automation in the field of construction. Hammad et.al [26] in their paper tried to build a predictive model for determining the duration of a steel fabrication using ML. The results turned out to be more accurate than the existing traditional practice. Some studies were done on estimating fluctuation costs such as liquidity, building services indexes by constructing classifier models such as SVM (Support vector diagram), Boltzmann machine (DBM) learning and back-

propagation neural networks [27]. Further, attempts have been made to use machine learning algorithms in occupancy predictions and Wi-Fi sensing in the buildings [28], and also it has been used to provide decision making output about the intensity of accidents occurred [29]. Other than these areas of construction, the use of ML has been very useful for temporary structures such as scaffolding. The ML adoption for the scaffolding works so far has been discussed below in section 2.5.

2.5. Scaffold related work

Scaffolding has become an irremovable concern to industries across petroleum, oil and gas, building, and infrastructure, in terms of less availability of labor and high cost, low productivity and delayed works [7]. It is believed that the in-depth analysis of scaffolding activities with a considerable amount of time spent, hardly have chances of success when compared to some other fields of study. The existing literature concerns the structural aspect of scaffold construction in specific scenarios. However, there are numerous researches that has proved its advantages over the machine learning algorithms finding solution for various construction related works. In addition, many scholastic researchers acclaimed that the planning methodologies hardly synchronizes with the dynamic scenarios and complexities faced by the construction industries. Campbell et.al [30] says that with the growing technology, computer simulations proved to be a solution for taking decisions during the planning of manufacturing processes since it has the capability of modelling the dynamic circumstances. Further, the author claims that simulation could also be useful in different stages of a project such as preconstruction planning phase, construction scheduling and post-construction [30].

The technology has grown wider and smarter to provide appropriate and accurate judgments for the complex scenarios in the construction field. The effective use of these resources can help in reducing potentially incorrect human judgments or decisions [6]. Although being introduced to a lot of advanced techniques and software, temporary works still lack essential

planning. Most of the temporary works, especially scaffolding, are generally not included in the main architectural and bid drawings [31]. Scaffolding being a significant temporary work often rely on the knowledge of engineers for estimation and planning. Since the front end planning of scaffolding mainly depends on individual experiences, there is a lack of firm scientific decisions and practical approaches for a proper estimation in terms of cost and manhours consumption [31]. There have been quite a few developments with respect to scaffold work. For instance, Feng et al. [31] in their paper describe constructing a safety model for scaffolding through BIM. A discrete firefly algorithm was used to find scaffolding scheduling and cost process for a modular construction environment. It was one of the attempts to address the issue of TCTP's (Time-Cost Trade-off Problems). Further, a multi-objective discrete firefly algorithm was created by Hou et.al [32] in which they used certain scaffolding work-related parameters such as the number of crews and equipment to achieve the objective of minimizing cost and time for scaffolding works in the projects. Cho et.al [33] in their paper tried to use machine learning algorithms (SVM-support vector machine) for predicting safe/unsafe situation of the scaffolding activity based on different loads carried. The model helped in deriving different safety scenarios such as safe situation, overturning, uneven settlement, or overloading conditions [33].In addition to safety and cost issues, there were few approaches where they tried recognizing the images automatically and learning the progress of the scaffold activities [34], constructing the BIM (building information modelling) based scaffolding framework to track the scaffold safety risks and finding standard hazards and solutions [31]. Hou et.al says [21] the main concerns of the researches so far has been about the safety and design related issues of scaffolding. Apart from these aspects, for the successful execution of any construction project, the other crucial issues that have been identified are optimization and planning of scaffold works. Addressing this issue has been one of the main concerns, especially in heavy industrial projects since they invest a lot of capital for machine

and labor. Kumar et.al [7] in their research says that, there is a difficulty in planning the industrial construction scaffolding because it operates mostly on as-needed basis .They claim that each scaffold manufacturing company have different approach in constructing a scaffold, which further leads to different cost estimates. Further, there might be a huge number of modifications done to a particular scaffold, since it is used for multiple trades. For instance, if a scaffold used for the first floor can be re used for the next floor just by extending few stairs; it is beneficial when compared to dismantling the whole thing and reconstructing the new one. These kind of scenarios make the scaffold planning more challenging. Further, he adds that these scenarios are labor intensive jobs and hence consume more time and cost. To overcome this issue, he suggests an advanced scaffold plan which anticipates such situations would be a beneficial to the companies. In his research paper, he made an attempt to carry out macro and micro level of estimation for scaffolding, considering various factors as inputs - trades of work, trade hours, the geometry of equipment, elevation of equipment, the weight of equipment etc. The output parameters involved scaffold manhours derived from the existing data. Similar to Chandan Kumar research, using WEKA analysis, researcher Wu, L. [19] came up with an estimation model for scaffolding by considering historical data of an industrial project. In these two research(Wu, L. 2013, Kumar et al. 2013), there were regression models created which gave a quantitative method for the project estimator and project manager to quickly come up with the scaffold manhours needed for future projects based on the information available at the start of a project. Due to insufficient scaffolding data, the results generated from the models did not yield good results. One of the recent studies of real time data of an actual liquefied natural gas (LNG) plant construction, address how scaffold productivity is affected by various factors such as scaffold types, dimension of the scaffolds and it discusses the obligations faced by the companies due to improper planning [35]. However, these approaches remarked as stepping stones to carry out the planning and estimation of scaffold research works.

To summarise, it can be said that poor planning leads to low productivity in the overall works and badly affects the performance of the project in terms of cost and time. In order to provide valuable solutions, this paper aims to develop a mathematical optimization model for manhours predictions in future use. In addition, this approach could be implemented in the planning stages of any heavy industrial construction projects, so that the project estimators get a standardized method to decide the manhours required for scaffold tasks. The next sections deals with the methodology adopted to build the predictive model.

CHAPTER 3 METHODOLOGY AND CASE STUDY

This chapter briefs about various types of data analysis and theories of machine learning techniques used in business and technology. This proposed research methodology describes step by step procedure of consolidating the collected scaffolding related data into an understandable format, cleaning the data errors by techniques such as outlier's removal, feature selection etc., visualising the data to learn the data parameters relationship with each other, understanding about different machine learning algorithms used for building the predictive model, and later to evaluate the model using error metrics for verification of the proposed model. Further, the results are briefly discussed.

3.1. Introduction to Data Analysis

Efficient data analysis helps to get necessary and valuable information from the existing records of data which provides significant benefits to the business and research fields. By carefully analysing the existing information, there are high chances of getting enhanced knowledge from past and present trends of any event or situation. In addition, the data driven decisions are very helpful for all sort of industries and business [36]. The general process of conducting data analysis includes defining of the problem statement, collecting necessary information, cleaning the data, and interpreting its uses for future purposes (Figure 9). Data can be further explored by the help of different statistical analysis and optimizations. There are different kinds of data analysis such as descriptive analysis, diagnostic analysis, predictive analysis and prescriptive analysis [36]. Each of them is discussed below.

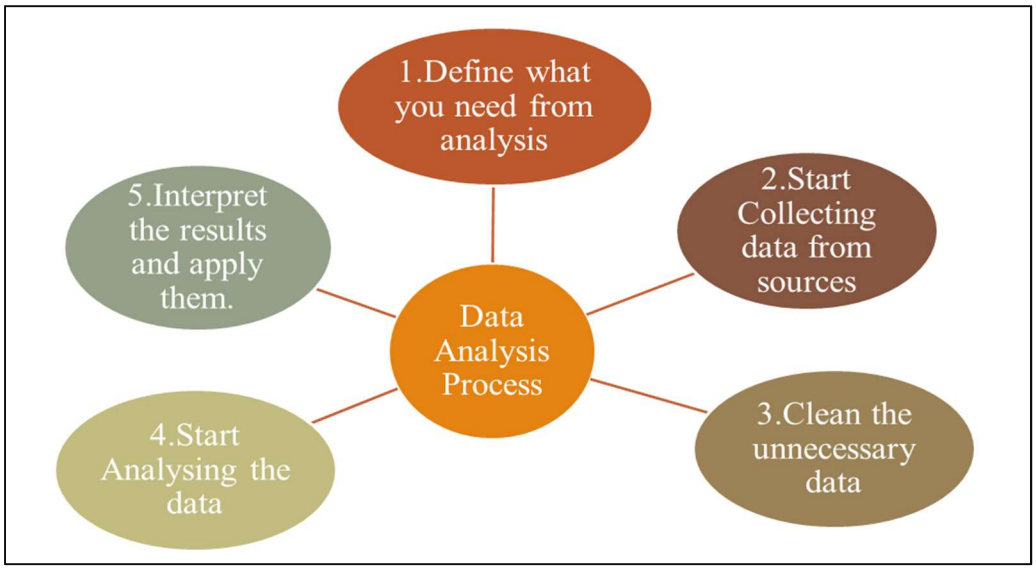


Figure 9 General Data Analysis process

3.1.1. Descriptive Analysis

Descriptive analysis yields the summary or useful statistics in an understandable format for the end users from the raw information. The analysis can be about detailed description of a thing, statement or an event that has already occurred. This type of analysis can be done verbally or statistically or sometimes both. Descriptive analysis would give better results based on individual experiences [36]. Examples of descriptive analysis includes sales overview of the companies, monthly revenue reports.

3.1.2. Diagnostic Analysis

Diagnostic Analysis are sometimes combined with the descriptive analysis. This type of analysis provides with more valuable set of information. The urge to gather more information makes the analysis a little harder to perform. In other words, this type of analysis tries to provide additional information about interconnections associated with an issue. For example, helping the customer to know what is the importance of choosing one product over the other [37].

3.1.3. Predictive Analysis

Predictive analysis is performed based on the trends held by the organisation's existing records. This analysis helps in forecasting the probability of an event occurring in future or estimating the time for the tasks to occur [36].There would be a lot of factors that the outcome would be dependent on. For instance, in the construction domain, particular accident risks can be predicted based on the existing activities responsible for the accidents [37]. Examples of analysis includes sales forecasting, risk assessment [37].The current research is a predictive analysis which helps in forecasting the scaffolding manhours for the upcoming industrial projects.

3.1.4. Prescriptive Analysis

Prescriptive analysis is the step by step explanation process of any situation. This analysis includes a systematic action plan that helps to obtain the required objective. It involves goals, values, policies and strategies to find a solution for the current problem or situation [36].The best example of prescriptive analysis is Artificial Intelligence [36].

3.2. Proposed methodology

The traditional practice for manhours estimation of each scaffolding task is based on scaffolding experts decision, which necessarily need not to be correct always and it is mostly subjective. As discussed earlier, the decision varies from person to person and also from company to company. There might be fluctuations such as the delay of work due to unfavorable temperature, skills of labors or other material related issues which alter the decisions, time and again. Also, few industrial companies claims that some sub-contractors assign the scaffolding manhours based on the trade of work to be carried (first preference would be given to piping), whereas few other companies uses the date of request raised (the early request is considered first) for the scaffolding activity. This often leads to impromptu decisions which would affect the time and cost of the projects adversely. Therefore, this research aims to identify all the

factors that are affecting the scaffolding manhours and find out a technical solution in terms of planning and estimation in order to avoid the excess manhours consumption from the scaffolding activities. A scientific and systematic approach to determine the scaffolding manhours in the planning phase of the project would avoid the ad-hoc way of handling the man power for the scaffolding activities, and also helps the supervisor to have a detailed plan for the associated scaffolding works in any given projects. Hence, predicting the scaffolding manhours using the existing data patterns is considered as the end output in this research because the early predictions during the planning phase would avoid schedule and cost overruns. The derived output (manhour spent on scaffold activities) is based on particular trades, work type, temperature, elevation, and other parameters. The process to achieve this objective, involves different stages which starts from collecting the historical data (scaffolding data) of the project (from a heavy industrial company), data preprocessing which includes cleaning the data and removing the outliers that disturb the data pattern, clustering the data into various groups, selecting the variables using feature selection process, training the data by different algorithms and yielding most accurate models for future predictions.The methodology consists of five main components:

i. ***Data collection***: In general, the scaffolding structures are assigned to any particular trade, location, or as a support for common access in the site area. To perform scaffolding activity, a request has to be raised by trade foreman to the scaffold supervisor of the project. These requests are sent for approval from the site engineer or senior person in the project. Once the requests are approved, the scaffolding structure is built, modified or dismantled in the field. The data related to scaffolding activity is then collected from the construction sites. There would be a scaffold coordinator, responsible for entering these requests and keeping track of the scaffolding progress on site.

ii. ***Data preprocessing***: After the data is collected, data preprocessing is conducted to - remove noisy and inconsistent data, merge the data from multiple data sources, clean and transform the data by removing the obvious outliers in the dataset [38]. The current preprocessing of the scaffolding data involves three steps: (i) consolidating the data received from multiple sources such as payroll, field data; (ii) using mathematical methods such as boxplot and scatter plots to identify outliers; (iii) removing the potential outliers by consulting the scaffolding crew and engineers.

iii. ***Data visualization*** – The preprocessed data is further visualized to understand the data in different perspectives, to know the trends of data workflow, also to observe the values in macro level and micro levels. Visualizing the data is done through creating correlation matrices to find the relationship of various factors which are affecting the scaffold manhours. The graphical charts are used to learn the relations between different parameters, proportions of different trades, night time work involved over the project duration, discipline of work, scaffold crew members required for scaffolding activities and other valuable insights.

iv. ***Predictive models*** –After understanding the data and its parameters , a predictive model is developed to forecast scaffolding manhours using various mathematical methods, and a comparison is conducted to find out which method is relatively optimal in terms of performance. Various machine learning algorithms are trained to build a model for future predictions.

v. ***Model Evaluation*** - The models built are regression type, which recommends different evaluation metrics such as RMSE, MAE and R squared value to find the better performance. These metrics helps to decide best suitable model for predicting scaffold manhours. Based on the existing data, by the aid of machine learning algorithms, it is possible to train and build the predictive models (using machine learning algorithms such

as regression and neural networks). The proposed model requires various input parameters that would be responsible for performing each scaffolding task. The built model would provide approximate manhours for scaffolding tasks in future, if the necessary information is fed. The process estimates and keep track of the manhours for scaffolding works and help during the decision-making process to plan efficiently. Each of the data analysis process is further discussed individually in the following sections. The Figure 10 represents the overview of the proposed methodology in this research.

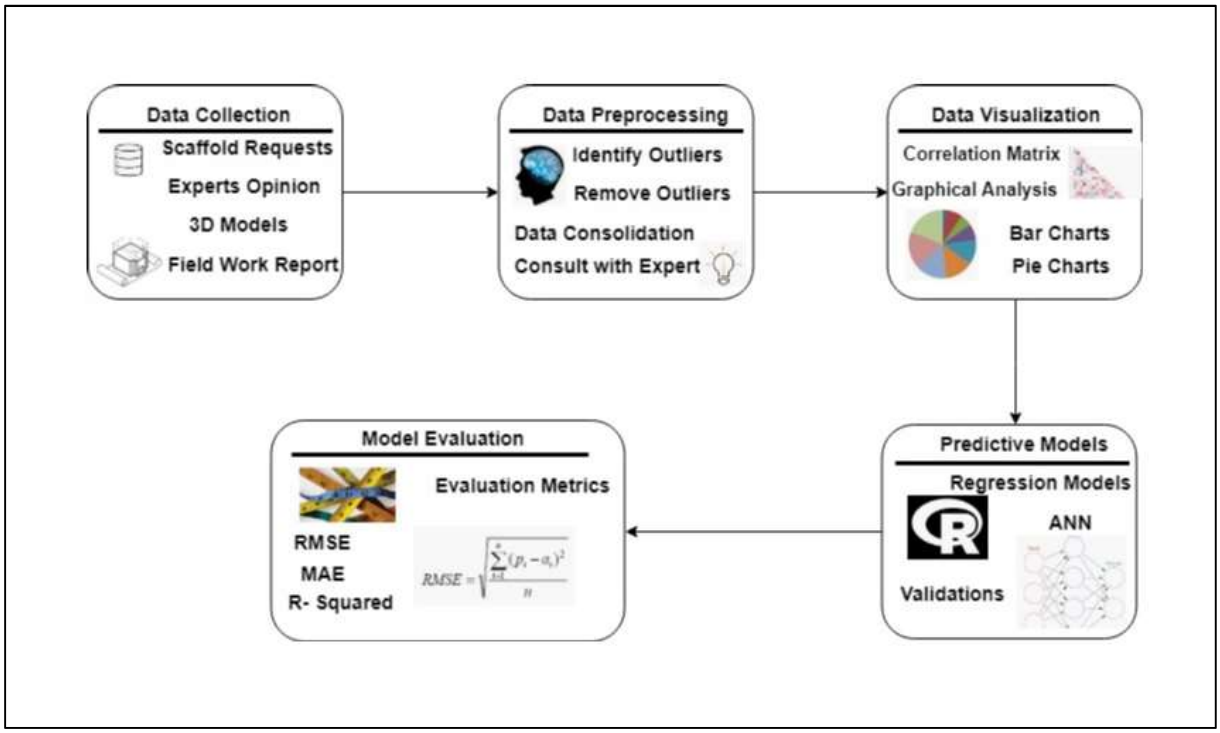


Figure 10 Framework of the proposed work

3.3. Data Collection

Data Collection is the initial step to conduct any data analysis process. Gathering all the necessary data would help in analysing the information and further use that source for getting the data driven and machine based efficient solutions. Collecting the relevant data from the company's information systems is the preliminary task to start with the data analysis. Numerous discussions were held with the related engineers and experts in the scaffolding field

to inquire and gather the necessary information. Generally, the scaffoldings are either built, modified or dismantled for different trades such as piping, civil, electrical, and mechanical. The scaffolding data involves many categorical and quantitative parameters. The quantitative parameters include elevation, temperature, weights of scaffolding materials and so on. The categorical parameters consist information such as the site area, discipline of work (build/modify/dismantle), and trade for which scaffolding is needed. In this research, the data required to conduct the scaffolding data analysis is gathered from three different sources.

i. ***Field collected scaffolding data:*** The general procedure for a scaffolding activity is carried out by, the trade foremen requesting to the scaffold supervisor or engineer for resources (labor and material), further the request is carefully assessed by the engineers. These requests are categorized into three types: (i) the new scaffold builds; (ii) modification of existing scaffolding; and (iii) dismantle of the scaffolding. Within each request, the details of the scaffolding activities are recorded. Each request, based on the entered information, has an estimated manhours for the corresponding scaffolding activities. After these activities are conducted, the information about the actually spent hours and used materials are retrieved back to the coordinator or engineer to enter into the system. Those data are tabulated in the spread sheets and other forms for the record maintenance.

ii. ***Location data:*** It involves work area volumes, location of work areas where scaffolding is built. The sources to track the location data are 3D models and the cloud based systems within the company.

iii. ***Progress data:*** It involves data from other project management systems, such as payrolls, materials list data etc. Such data consists project completion percentage, actual man hours, additional hours, temperature, night time ratio, apprentice ratio, work shifts, quantity of materials used for scaffolding and its weights.

In this research, a heavy industrial project involving multiple trades such as piping, electrical, and civil is considered as the case study to analyse the scaffolding activities. The scaffolding activities carried out for different trades in that particular project has been gathered. The data is further consolidated to a structured pattern by removing outliers, selecting the important parameters and clustering the dataset. To get the effective end results for any data mining process, it is necessary to get the complete data before performing any technical analysis. The particular project considered as case study in this research was completed in a span of 30 months. The scaffold requests details collected for each scaffold task were in separate excel spread sheets. Further, tabulating and extracting the information from multiple tables to one single table were performed using R programming language. All data from the sources were eventually consolidated as a dataset for further data processing. The final single dataset which was considered for this research had around 15000 rows (data points). Each row of information represented a scaffold request and had a unique Task ID. The raw data of the project considered in this research, had 17 parameters. Each of the parameters are briefed in Table 1. There were 5 categorical variables and quantitative parameters. Table 2 represents the sample data set of the project considered for case study. Further, after data collection there is a need to process the data into clean and understandable format. Hence, various statistical methods and logical implementations were used. Each of them are briefed in the next section.

Table 1 Parameters related to scaffolding data

Categorical Parameters	*Quantitative Parameters*
Specific Areas: *Construction areas where scaffolding activities are conducted*	**Working Date(dd/mm/yyyy)** *The start and completion date of the scaffolding task*
Work Classification: *New build (erection), modification, or dismantle*	**Actual Manhours (hours)** *The tracked actual spent manhours of the scaffolding tasks*
Scaffold Type: *Typical types include: Platform deck, tower, barricade etc.*	**Total Added/Dismantled Weight(lbs)** *Weight of the scaffolding materials added or removed from the scaffolding structures.*
Discipline: *Discipline of trade work that scaffolding are built/modified for. E.g. Civil, structure, mechanical etc.*	**Elevation(ft.)** *At which height scaffolding activities are conducted relative to ground elevation*
Shift : *Day shifts or night shifts of work*	**Temperature(^{0} C)** *At which temperature the scaffolding tasks were performed on site.*
	Delay hours (hours) *If there were any delays due to labor or equipment unavailability.*
	Night Shift Ratio (%) *Percentage of work done in the night for each scaffold task.*
	Apprenticeship Ratio (%) *Percentage of apprentice labor worked for each scaffold task.*
	Man count(No.s) *Number of labor required to perform each scaffold task.*
	Aluminium Percentage (%) *Percentage of aluminium in the scaffold materials used for each scaffold task.*
	Workable Area (Sqft) *Available work space area to perform each scaffold task.*

Table 2 Sample data set for scaffolding work

Task ID	Working Date	Man Count	Delay Hours	Discipline	Temperature (degree celcius)	Work Classification	Scaffold Type	Shift	CWAs	Elevation (metres)	Apprenticeship Ratio	Night Time Ratio	Aluminum Percentage	Weight of the scaffold (Lbs)	Number of major pieces	Workable Area (Sqft)	Manhours (hrs)
1	02-10-2016	10	0	Piping	12	Erection	Cantilever	Day	CWA -1	19.2	0.40	0	0.00	8600.3	400	680.5	120
2	03-10-2016	8	3	Electrical	11	Erection	Bridge	Day	CWA -2	16.5	0.54	0	0.21	3564.9	200	469.4	86
3	04-10-2016	5	2	Civil	9	Modification	Hoarding	Day	CWA -3	13.2	0.15	0	0.13	2000	100	256	61
4	05-10-2016	4	1	Mechanical	3	Modification	Tower	Day	CWA -4	12.1	0.56	0	0.51	100	8	123	8
5	06-10-2016	6	0.5	General Management	7	Erection	Tower	Night	CWA -5	18.1	0.38	0.3	0.69	152.6	60	230.12	7
6	07-10-2016	4	0	Structural	6	Erection	Hoarding	Day	CWA -6	13.1	0.22	0	0.43	700	15	236.5	73
7	08-10-2016	4	0	Sub Contractors	4	Erection	Working Deck	Day	CWA -7	14.2	0.44	0	0.12	2500	31	278.6	64
8	09-10-2016	4	2.5	Civil	6.5	Erection	Hoarding	Day	CWA -8	16.8	0.40	0	0.23	980.3	86	33.3	71
9	10-10-2016	6	0	Mechanical	2.1	Erection	Dance floor	Day	CWA -9	17.8	0.80	0	0.00	610.2	32	123.12	54
10	11-10-2016	6	3	General Management	3.2	Erection	Working Deck	Day	CWA -10	13.5	0.60	0	0.30	1560.4	40	156.6	31
11	12-10-2016	6	0	Structural	1.9	modification	Tower	Night	CWA -11	15.5	0.90	0.4	0.30	1966.5	120	356.5	45
12	13-10-2016	7	0	Sub Contractors	3.8	Erection	Barricade	Day	CWA -12	17.4	0.60	0	0.56	169.86	50	465	48
13	14-10-2016	7	0	Piping	-0.8	Modification	Tower	Day	CWA -13	24.7	0.60	0	0.34	6666.5	180	560	186
14	15-10-2016	10	0	Electrical	-1.6	Modification	Barricade	Day	CWA -14	12.3	0.60	0	0.61	7014.7	420	2400.6	95
15	16-10-2016	8	1	Piping	2.4	Modification	Working Deck	Night	CWA -15	14.2	0.80	0.2	0.30	1635	186	1900.655	90
16	17-10-2016	10	2	Electrical	0.9	Dismantle	Tower	Day	CWA -16	17.1	0.60	0	0.31	2800.6	375	245.6	79

3.4. Data Preprocessing

According to the data analysis process, the next step after data collection is data pre-processing. Generally the data collected from the large databases will have errors. These errors might be in different forms - (i) incomplete data which has certain missing attributes values, contains aggregate data. (ii) Noisy data which has errors and outliers values that deviates from the larger pattern of data. (iii) Inconsistent data which has discrepancies in the codes, negative or incorrect values. These kinds of errors are quite common in real world databases due to various reasons such as manual entries, multiple people handling the data, computer errors at time, technical limitations while transferring or merging data, incorrect format of entering the parameters and so on[39]. The data which has to be analysed using data mining techniques should be in a clean and consolidated state. The processing of data involves multiple steps such as data cleaning, data integration, data transformation and data reduction. Data cleaning involves filling up the missing values, identifying the outliers and removing them using statistical methods. Data integration is a process of merging data from multiple databases and data transforming is a process of normalising the data to have the uniformity in the distribution of values. Further, data reduction is removing unrelated parameters (by correlation and other feature selection methods) which does not add value to end result of data mining process [39]. In this research, the initial step of data pre-processing, that is data cleaning is done by removing the uncertainties such as missing/incorrect values, duplicated data entries, and removing the potential outliers with the help of statistical methods and expert's opinion in the company. After the scaffolding work is executed, the manual entry related to the works are tracked and recorded by the particular scaffold foremen or the coordinator. The tracked data might have some errors, such as lack of information about a particular scaffold request or repetitions of the same requests. These values/rows were filled up by understanding job descriptions for each requests and particular scaffold supervisor for the project was contacted for the assistance in

removing the rows that had missing values. Due to the errors resulting from the human components in the data collection processes, there are high chances of having outliers in the data that can affect the results of data analysis process.

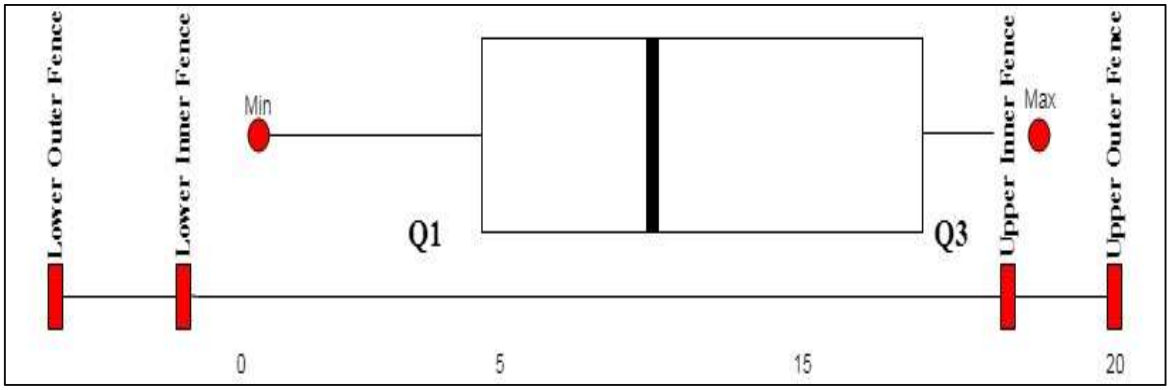
The data points in the dataset which do not lie in the general behaviour of the model are referred as outliers. They disturb the distribution pattern of any data set [39]. There are various methods of handling the outliers such as least square fitting (regression), standard deviation, interquartile range (IQR), and cook's distance and so on. In addition, there are visual methods such as scatter plots, boxplots and histograms help to analyze the behaviour of data and identify outliers. However, there is no standard rule that is followed to remove the outliers. In past research, data analysts have tried different approaches to remove outliers from the databases. For example, Negri et al [40] introduced an artificial neural networks (ANNs) based method to estimate the consumption of electrical materials in early stages of construction projects. In their research, higher root-mean-square-error (RMSE) is used as an indicator for outliers (to be specific, 10% of the attributes with the highest RMSE is removed from the dataset to protect the model). Huang et al. [41] have proposed a novel strategy of fusing available redundant measurements of cooling load of multiple-chiller plants to reduce measurement uncertainties, where the Moffat distance is used in consistency checking to remove the outliers. Another example is cluster method for identifying text similarities in construction documents that cluster size is used to identify outliers [42]. Although the past research has addressed outliers from different statistical methods, the outlier removal process remains similar. The first step will be identifying the data points in the dataset that are far from the overall distribution pattern (e.g. by mathematic quantification such as percentage, distance or data sizes); and the next step is to decide the potential outliers by justifying and confirming with the experts who completely understand the data. In this research, an iterative method IQR (Inter Quartile Range) is adopted to identify potential outliers. Further, the scaffolding expert's opinion was considered for

removing outliers. Figure 11 shows the general representation of IQR ranges through the boxplot. To illustrate IQR, the box plot defines lower quartile as the 25th percentile, median as the middle point and upper quartile as the 75th percentile. The difference between the upper quartile (Q3) and the lower quartile (Q1), which spreads over half of the data is referred as IQR.

$$IQR = Q3 - Q1 \qquad \text{……Eq (1)}$$

A point beyond the inner fence on either side is considered a mild outlier. A point beyond the outer fence is considered an extreme outlier. A potential outlier is a data point that is 1.5 times the IQ range from the edge of the box [39]. IQR method is similar to the standard deviation; however, IQR range is advantageous because they are not affected by the extreme values of data points [43].

Figure 11 Boxplot showing IQR ranges



As mentioned earlier, there are no written rules to remove outliers. However, in this research a set of rules were defined to remove outliers. Since, the scaffolding man hours is the desired output, removing the out of range values from the manhours could be the suitable option. But, in some cases, if the manhours consumption is more may be the work completed in terms of weights or volume would also be more. In this way, it would be inappropriate to remove outliers just based on manhours. Hence, a new parameter, which determines the hourly productivity of the scaffolding works was introduced. The productivity of any construction works is generally

measured in terms of duration taken to complete any given task. In this research, productivity is defined as weight of materials carried in each manhour.

$$\text{Productivity} = \frac{\text{Total Weight of materials used for each scaffold task (lbs)}}{\text{Total Manhours to complete each task (hrs)}} \quad \ldots \text{Eq (2)}$$

There were two methods adopted to identify outliers – *(i) Statistical method (IQR) and (ii) Company official's rule (based on project knowledge)*. IQR range was established for the productivity in the data set. The default value of IQR (values less than and more than 1.5*IQR) resulted in removing large amount of data from the data set (almost removal of 3000 rows/data points). Further, when the results were discussed with scaffold officials, removing necessary data just based on analysis of IQR was not accepted. However, it was suggested to try removing the data points with the help of different combinations such as removing data points below 25% and after 75% range of data set. For instance, removing the 5 % of data which was not in the range compared to the rest 95% of the data set. The different probabilities combination such as 20%-80%, 15%-85%, 10%-90%, 5%- 95% could be tried to choose the data set in such a way that there is no unnecessary loss of data. It is worth mentioning that, in real-time data, it is good to seek opinion of experts who know about the data well and assist about the necessity of removing the outliers. In this case study, after identifying the outliers through IQR range, the second approach was to consult with the company's professional scaffold practitioners. A new set of rules was made by them for removing the potential outliers. The rules set by experts knowledge, removed the data points which fall in the following criteria.

i. The data rows which doesn't have manhours value and the ones less than 5.
ii. The rows which had less than 20 lbs weight.
iii. Productivity value below 5 lbs/hr and above 200 lbs/hr.

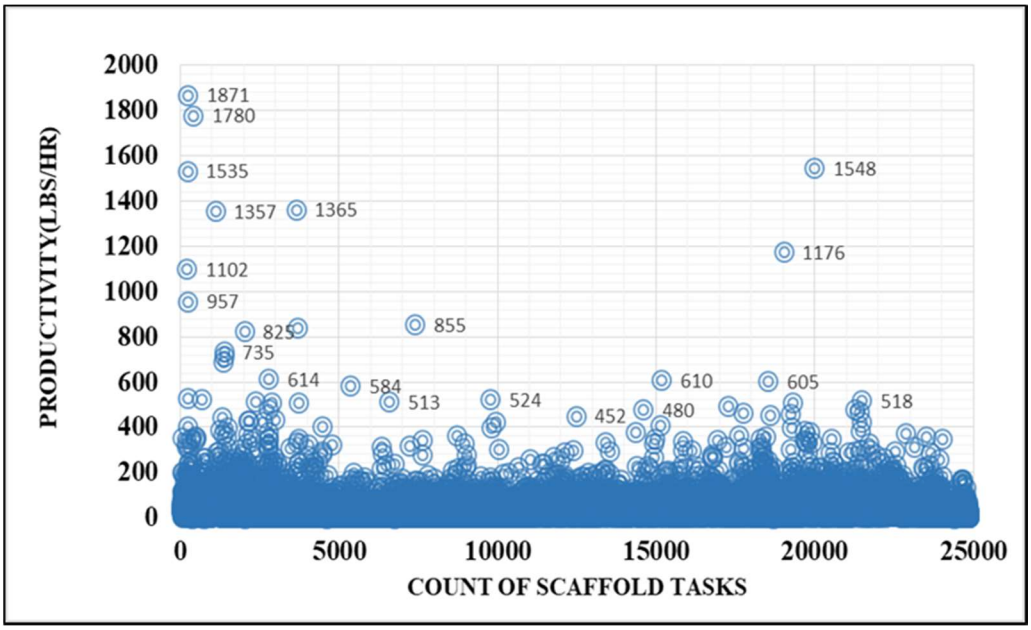
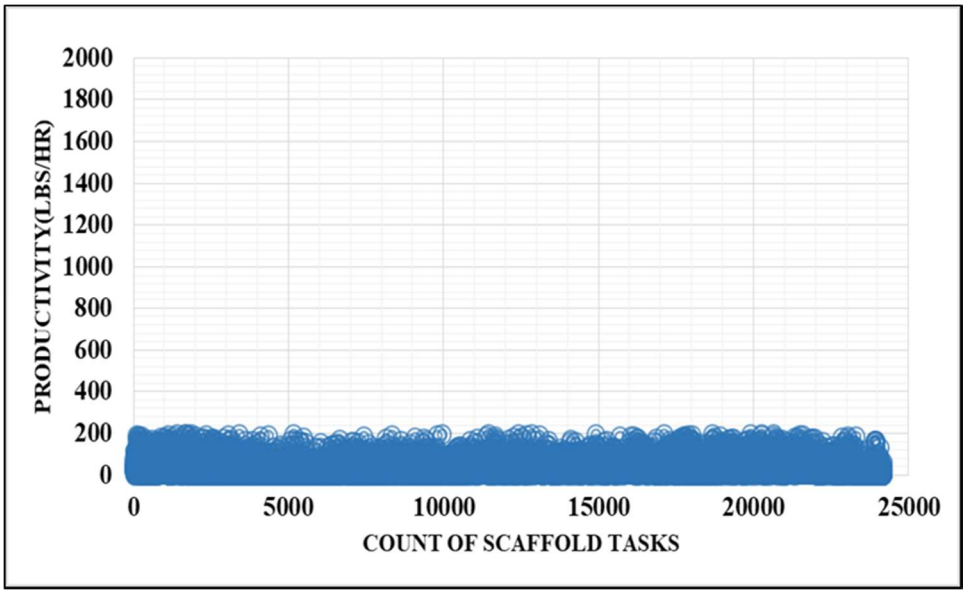
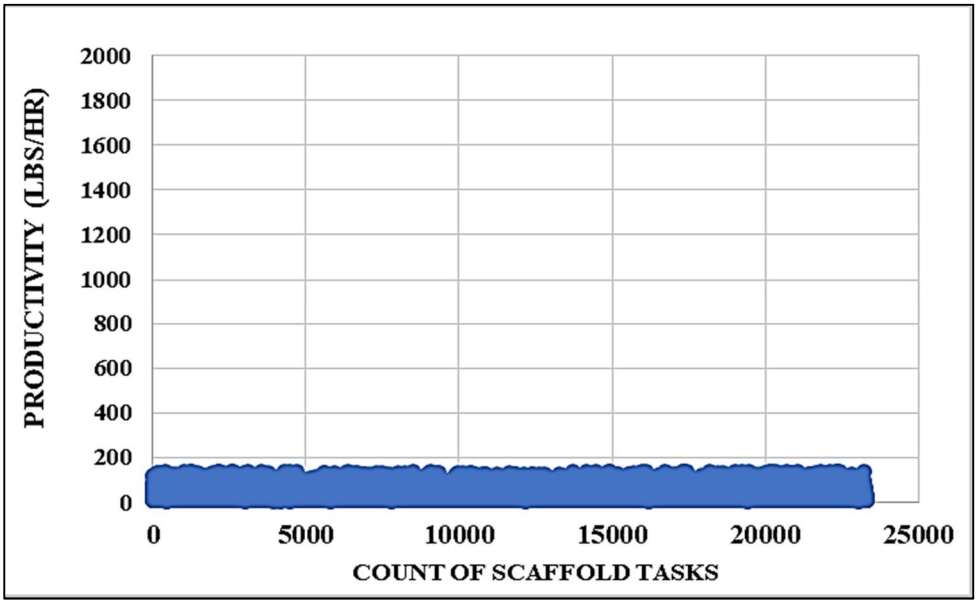


Figure 12 Data set before removing outliers

The Figure 12 above represents the scatter plot of productivity of scaffolding tasks accomplished during the project life cycle. Scatter plots after removing the outliers have a uniform pattern from both company's set of rules represented in Figure 13*(i)* and statistical range by IQR represented in Figure 13*(ii)*. Further, the different data sets obtained from each IQR range was trained to build the model and compared with the model developed from the dataset obtained by applying company's rule. Based on the careful observation, probability range of 5-95% was chosen because the number of data points in this range almost matched the number of data points of company's rule. Another reason to choose 5-95% range is, the data should not be unnecessarily wasted by removing outliers in extreme ranges. However, all the other probability ranges are further to be checked in the later stages.



(i) Data set after removing the outliers (company's rule)



(ii) Data set after removing the outliers (IQR rule [5-95%])

Figure 13 Scatter plot of Manhours after removal of Outliers.

After removing the outliers the necessary step is to remove the parameters which are not impacting the output variable. Along with the required information, the other project centric details from the data, which does not add value for future predictions were carefully analysed and removed by seeking the scaffold experts' advice. The scaffolding data set had 5 parameters of categorical type - work area, discipline of the work, work classification, shift and scaffolding type. The 'work areas' represented the location of work for only particular project, hence those

data were not useful for the further analysis. Adding on, the 'shift' parameter was also considered insignificant for the analysis, since a quantitative parameter 'night time ratio' addressed the timing of shift time (day/night) in more specific way by mentioning the exact percentage of work done during the day and night shift. Hence, the data set comes down to 3(out of 5) main categorical parameters namely work classification, scaffold type and discipline of the work. The micro level distribution of these parameters are shown below in the Figure 14.The distribution of man hours and productivity for all these parameters were visualised further with the help of various graphs.

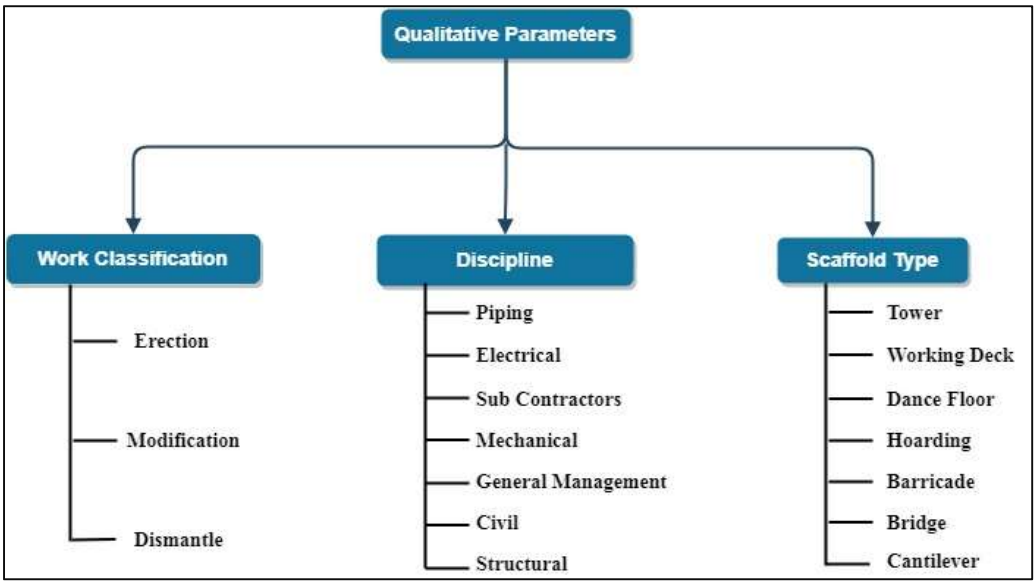


Figure 14 Micro level distribution of qualitative parameters

Studying the data using spreadsheets or reports would consume more time for larger datasets, so analysing and visualising the data represented in the form of charts and graphs would be more convenient to know more insights about the data. Hence, the data visualising is an essential step in the data analysis process.

3.5. Data Visualization

Data visualization provides quick and easy way of conveying information from any particular set of data. It helps in identifying the parameters which of them requires more importance and the ones which can be ignored in the data sets. Further, data visualization is not only about standard charts and graphs; it assists on patterns, trends, and correlations of the data by providing heat maps, sparks lines, regression lines, infographics and lot more, making it aesthetically richer than information visualizations [44].Data visualisation helps in finding relationships and comparison between different parameters, also the distribution and compositions of parameters in the data sets. The Figure 15 below shows how the visualisation can be used in different form of the graphs and charts. Each type of graphs were further generated for the case study to analyse the data visually.

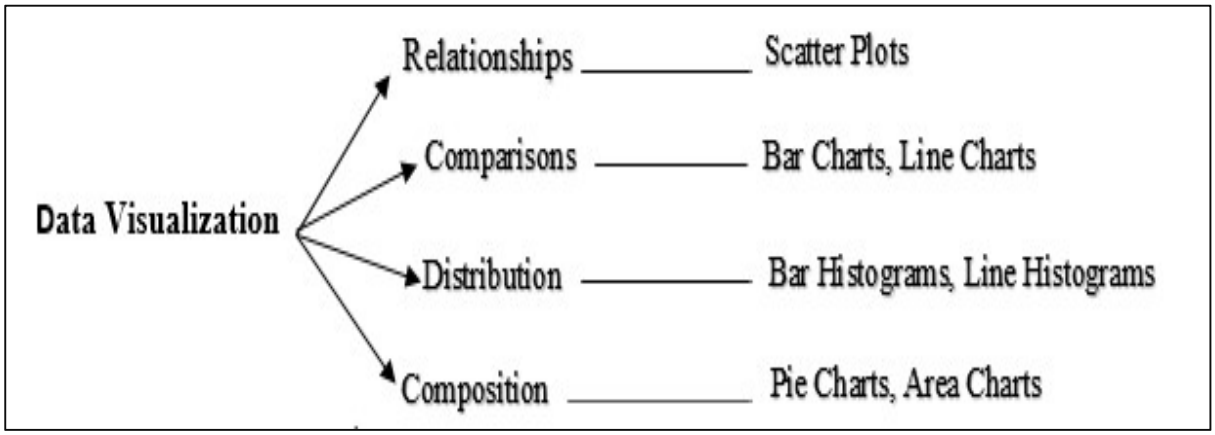


Figure 15 Components of Data visualisation

As mentioned earlier, the scaffolding data had multiple numerical and categorical data, each categorical data further had different sub categories. Henceforth, learning the data by visualising was a feasible decision. By the help of pie charts, it was able to analyse the percentage of manhours consumption by each of the sub categories involved in work classification. In Figure 16, as we can notice that erection and modification were highest manhours consumers. The reason is quite obvious that dismantling is easier and do take less

time. Further, the company and scaffold experts claims that the dismantle hours of the scaffold should be generally one third of the time required to build (erection). In this way, by observing the percentages in the graph, it can be said that the data makes sense in terms of work classification. In Figure 17, we can notice the comparison between the productivity ranges of each work classification. The productivity of modification is almost low as erection, the reason is adding and removing weights would consume more time .It is worth mentioning that, when the data parameters are visualised, the constant discussions with related supervisors will be mutually beneficial to company for getting the deep insights of the work and it also helps the data researcher /analyst to head in the right direction.

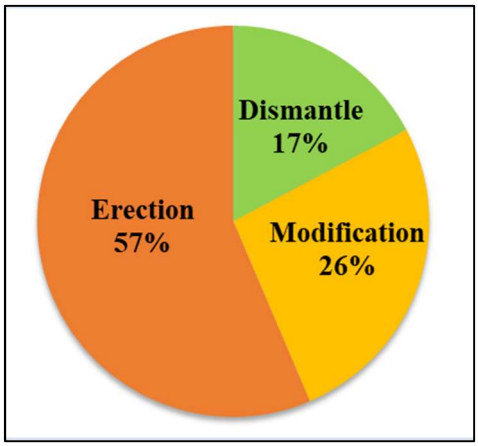


Figure 16 Scaffold manhours in terms of work classification

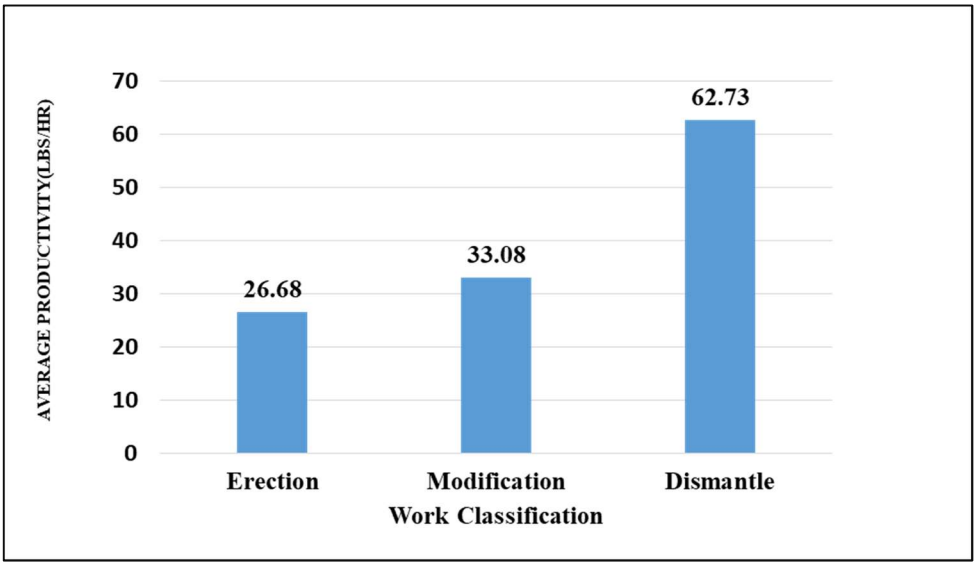


Figure 17 Productivity in terms of work classification

The consumption of manhours varies for different trades/disciplines. The scaffolding works built for general access to the labor works consumes less manhours due to ease of construction. However, the scaffold built for piping and other heavy works should be strong enough to withstand the loads carried on. Hence, those disciplines might consume more man hours and their productivity is usually low. In terms of discipline or the trade for this particular study, it is evident that Piping, Electrical and Sub Contractors (fire proofing and insulation) have the highest percentage of scaffolding manhours as represented in Figure 18 and the least productivity observed in Figure 19 .The other disciplines hardly impact the scaffolding manhours. This type of visualisation helps in suggesting the scaffold authorities, to know which parameters are spending more man hours, so that it guides them in exploring more about that particular parameters. Further, Figure 20 shows how the combined chart of work classification and discipline provides good visual analysing of the facts that where exactly the man hours are spent on each trades and each type of work.

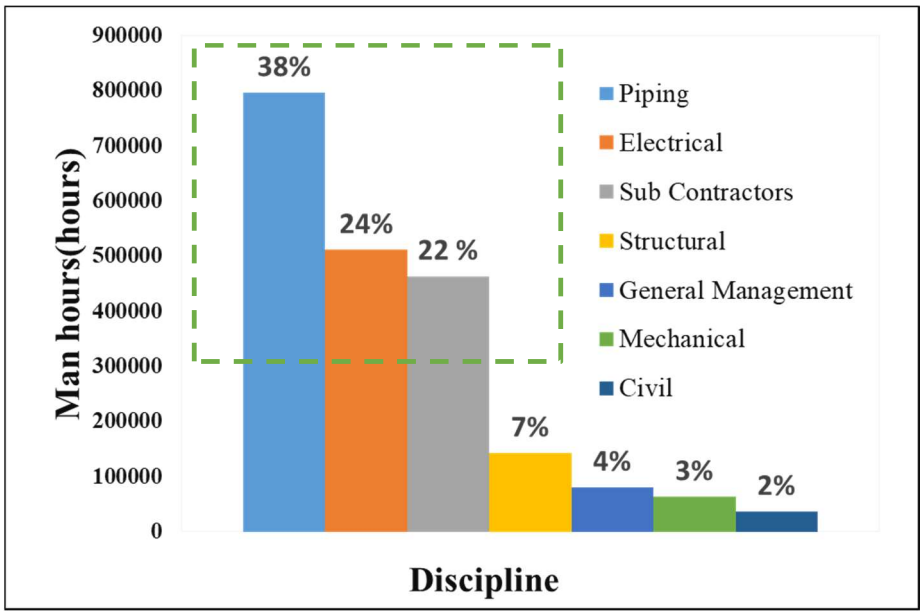


Figure 18 Scaffold manhours in terms of discipline

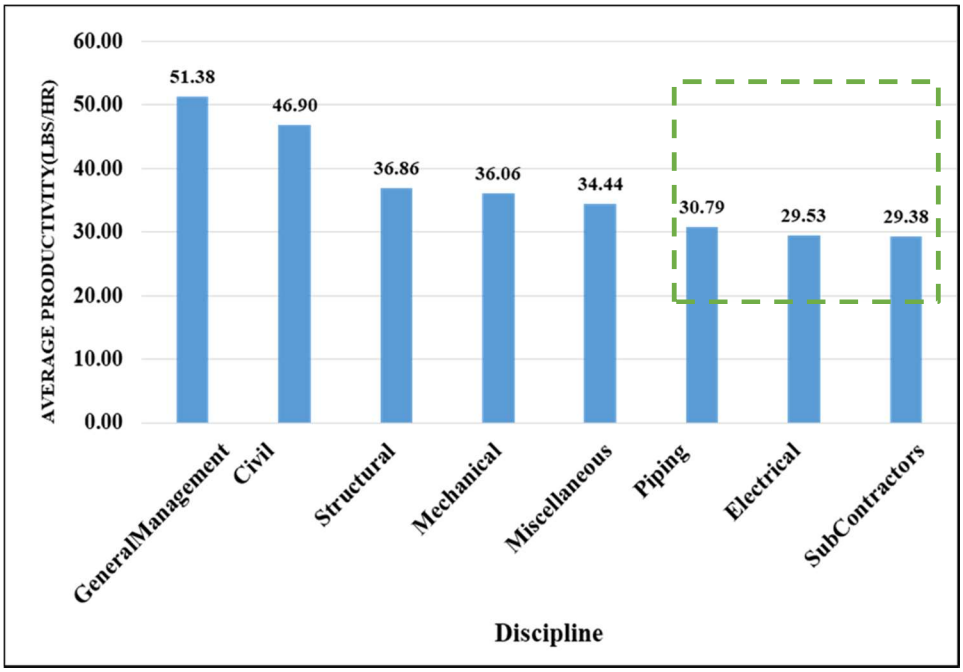


Figure 19 Productivity in terms of discipline

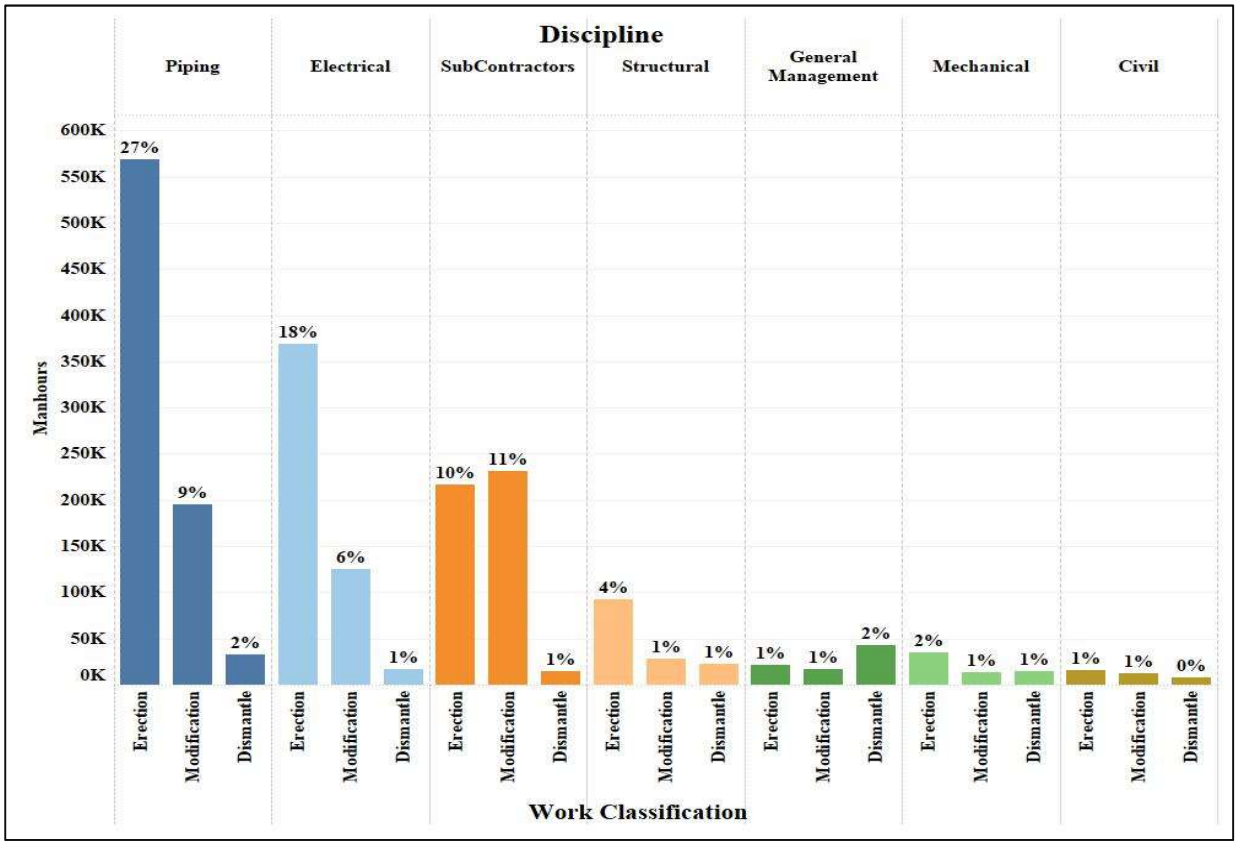


Figure 20 Distribution of man hours combined in both work classification and discipline

The type of scaffold for sure would be an important parameter when multiple types of scaffolding are involved in the project since they differ from each other in various aspects such

as weights, ease of construction, skills of labors to construct that scaffold and so on. On the other hand, some scaffold experts [4] also claim that any type of scaffold has more or less the same type of materials used to build. Hence, it's not a major concern for industrial sites. However, as per the data set information in this research, the type of scaffold used abundantly in the project was tower scaffold (almost 70%). Productivity is consistent in most of scaffold types except a fluctuation in bridge, hanger and barricade type of scaffold, which has less manhour consumption and they hardly had any impact on overall man hours for scaffolding activity. In the Table 3 we can notice that, there is a clear indication that Tower Scaffold has the highest percent of man hour's consumption. In the future analysis, this factor can be potentially ignored since there is no strong influence of the scaffold type on man hours.

Table 3 Scaffold type in terms of man hours and productivity.

Scaffold Type	*Average Productivity*	*Percentage of scaffold man hours*
Tower	32.71	69%
Dance Floor	30.65	8%
Cantilever	33.77	6%
Working Deck	38.45	5%
Hanger	22.79	3%
Hoarding	36.25	3%
Barricade	25.50	1%
Bridge	27.05	1%
Shelter	42.12	1%

Other than visualising the parameters through graphs and charts, for a better understanding; if the parameters are analysed along with the life cycle of the project it would give better insights. In addition, it would also give a clear picture of ups and downs of the parameters in the project such as, at which point excess scaffold man hours are spent and where productivity has been affected. For example, the Figure 21 below shows how the manhours is distributed along the

time period of the project. It can be interpreted that the manhours for scaffold is slow in the initial stage, then it got better towards the middle of the project .Eventually, at the end stage in order to complete the project within due dates, there was a rise in the scaffold man hours.

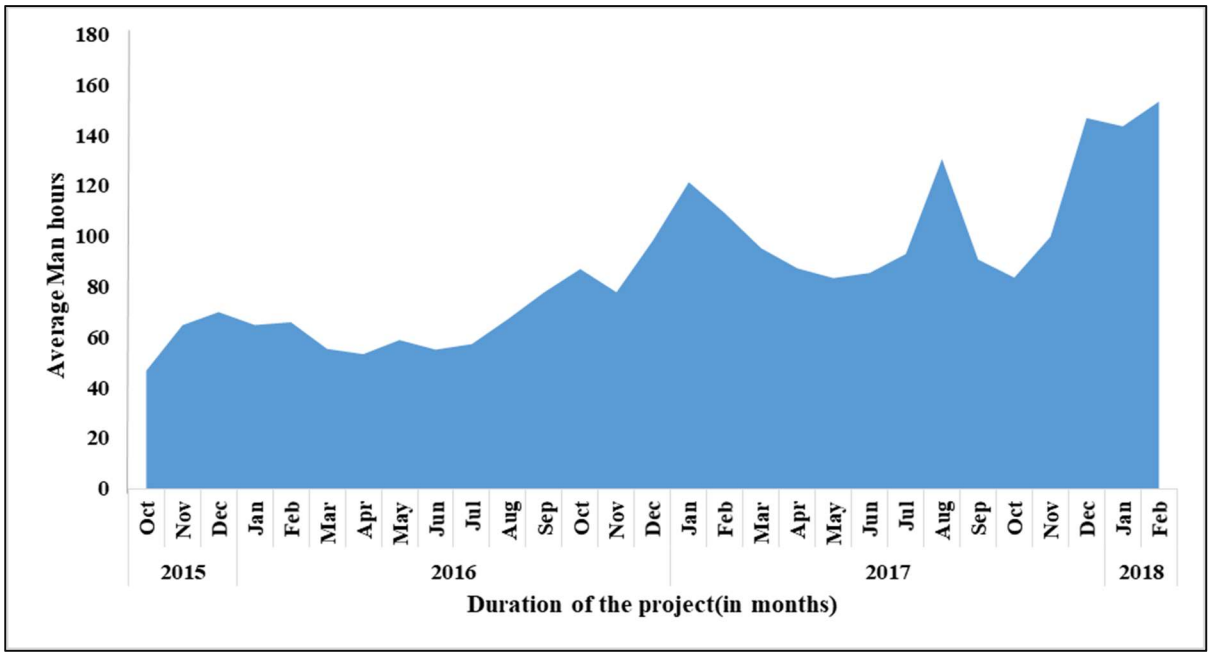


Figure 21 Average manhour distribution along the time period of the project.

There were additional graphs generated along with the lifecycle of the project to explore data in all possible ways. For instance, measuring the manhours of work along the life cycle of the project based on each work classification in Figure 22.It indicates that the initial stages hardly had more hours and with time, the manhours consumption levelled up at the finishing stages. It can also be noticed that erection has been the highest consumer during the midway of the project. However, the dismantle hours had a little raise in the end stage.

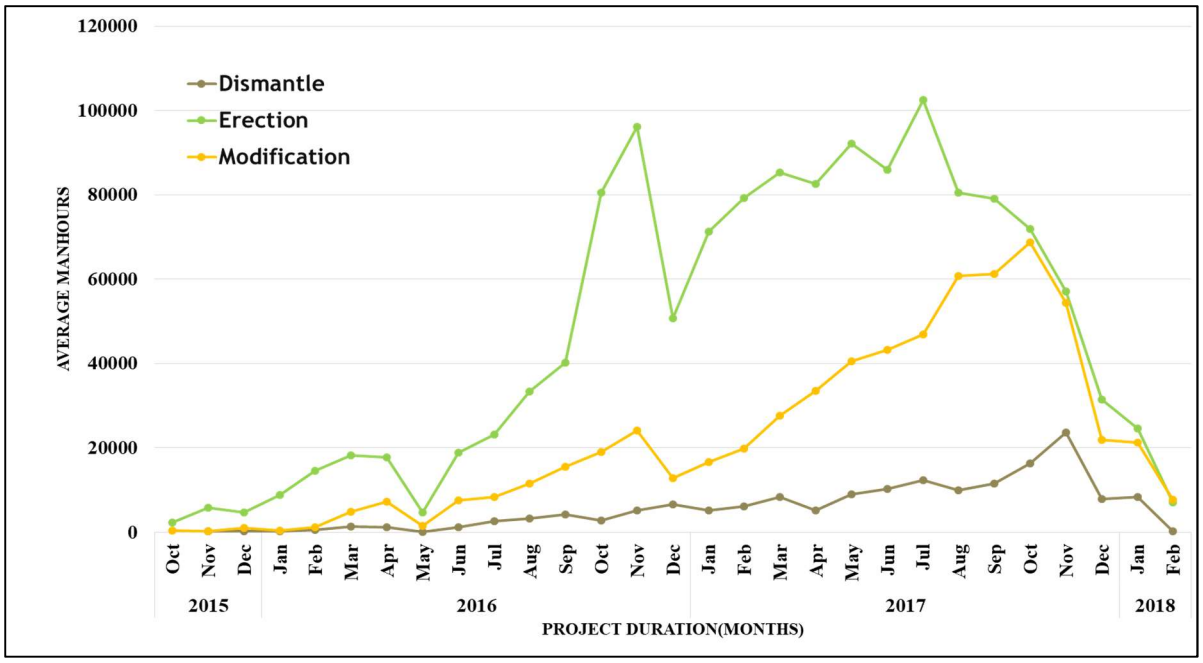


Figure 22 Consumption of manhours in terms of work classifications along time period

The quantitative parameters can also be analysed visually. For example, the productivity of work might be better in moderate temperatures compared to the cold temperatures. Figure 23 represents the temperature at which more manhours are spent, it might indicate that the man power is utilised well and the work have been progressed more in the temperature range 10-20 degree Celsius. However, the productivity chart represented in Figure 24 clearly conveys that after minus 10 degrees of temperature, the productivity is more or less the same. This temperature range above minus 10 degree can be considered as an ideal temperature for performing scaffolding work.

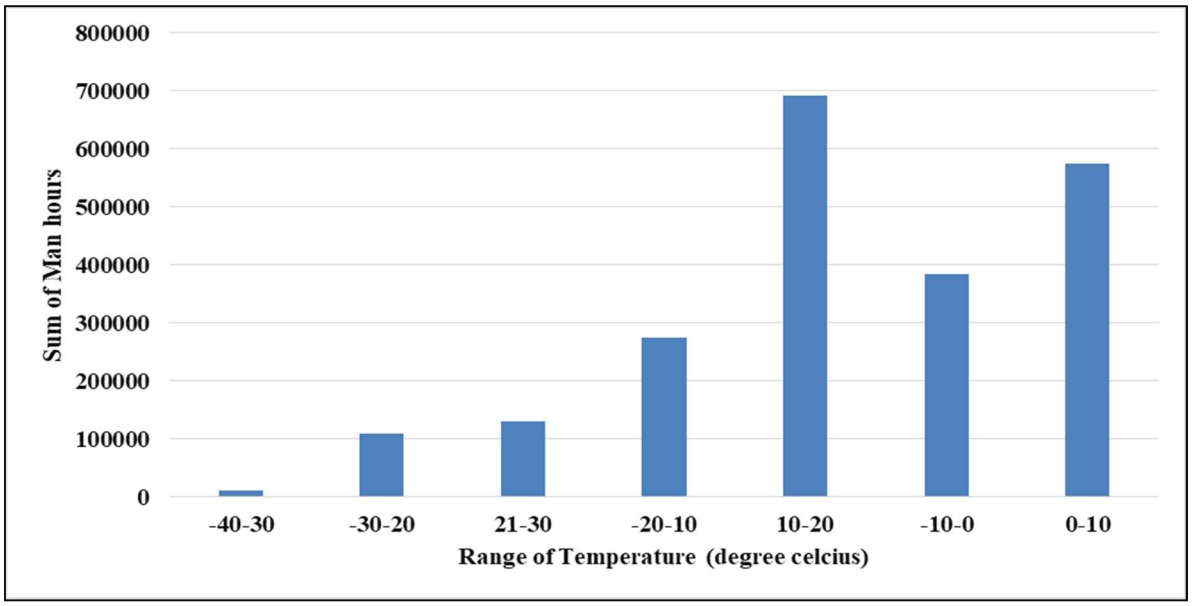


Figure 23 Manhours consumption at different temperatures

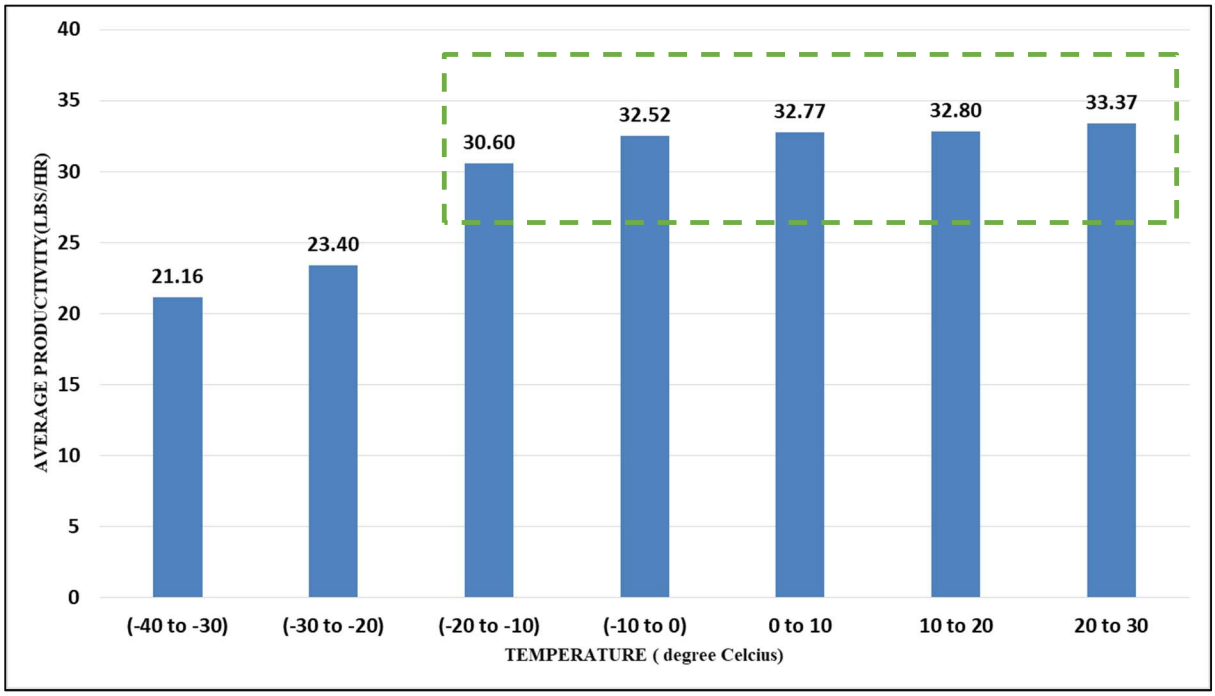


Figure 24 Average Productivity for different temperature range

Many other observations were made to provide the company a complete idea of the workflow and other related information regarding the project. For instance, the visualisation of delay hours along the time period of the project would help in taking care of the delay reason in future projects. Further, in order to select the parameters after removing the outliers and visualising the different factors impacting the man hours, many interpretations were made. The next step

is to find the interdependency of the parameters defined in the dataset. The level of interdependency can help in figuring out whether the parameters can be further used for building the predictive model or not. While building any machine learning algorithms, they use feature selection process to identify the dependency of output parameters on various input parameters. Feature selection methods are further discussed in brief.

Feature Selection:

Feature selection can be described has a process of identifying and removing irrelevant and redundant information from any data set [45].There are three different methods of identifying the variables impacting the output in predictive model building- filter methods , wrapper methods, and embedded methods. In order to reduce the data efficiently, a feature selection method can be used in the data pre-processing stage. This process helps in identifying accurate data models [46].

i. ***Filter methods*** – It uses statistical tests and measures, then assign a scoring for each feature or variable. The variables are ranked and based on the score they are either selected or removed from the dataset. Examples of filter based methods are Chi squared test, information gain and correlation coefficient scores [46].The process is represented in Figure 25.

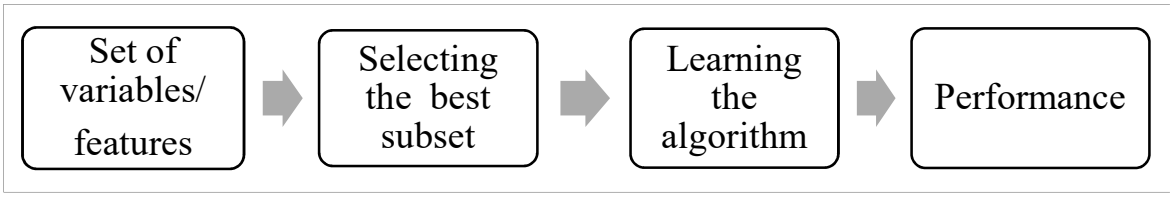


Figure 25 General process of filter method feature processing

ii. ***Wrapper methods*** – A wrapper method will choose a particular feature subset based on how effectively a modelling algorithm performs, this is considered as a black box evaluator. Thus, a classifier performance is chosen to evaluate a subset for classification tasks and similarly a cluster algorithm's performance is considered to evaluate a subset for clustering

through wrappers [46]. Examples of wrapper methods are random hill-climbing algorithm, Boruta algorithm, forward and backward propagation. The process is represented in Figure 26.

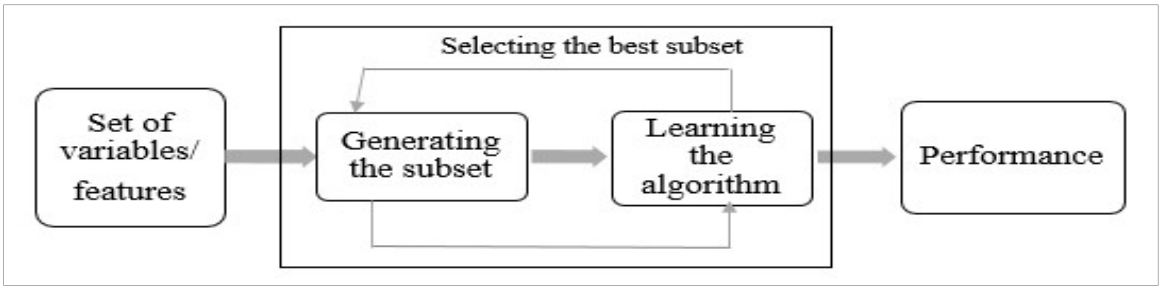


Figure 26 General process of wrapper method feature processing

iii. ***Embedded methods–*** These methods as the name goes are embedded in the algorithm either as its normal or extended functionality and perform the feature selection processes during modelling algorithm's execution [46]. Examples of embedded methods are stepwise regression, regularized trees. The process is represented in Figure 27.

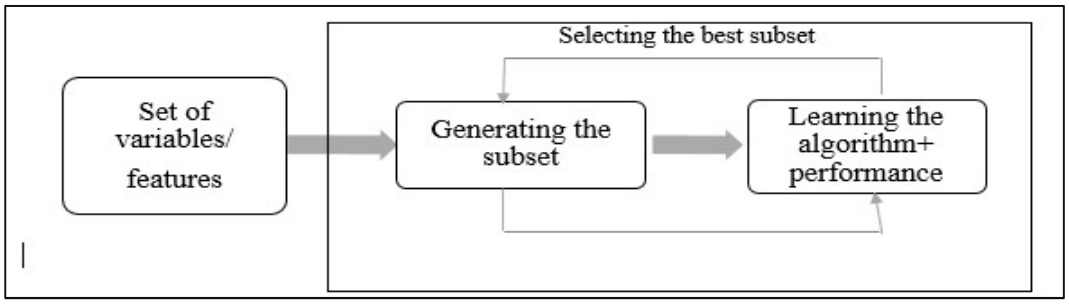


Figure 27 General process of embedded method feature processing

In this particular research, the adopted feature selection processes are from all the three different methods and the results are compared to choose the variables to build model. Statistical experts do suggest that initially linear predictors should be used. Further, they suggest different subsets has to be tried to cross check the performance [46]. Feature subset selection process allows the machine learning algorithms to run efficiently and improve the performance by eradicating any unnecessary and repeated information from the data. This can be seen in the form of either accuracy in future classification or a better representation of the targeted result [47].The feature selection methods adopted to select the variables are

iv. Correlation matrix (Pearson's coefficient)

v. Random forest method

vi. Step Wise Regression

Correlation matrix (Pearson's coefficient)

A correlation can be defined as a statistical technique as to know how two variables are strongly related to each other [48]. A correlation matrix is a table that represents the correlation coefficients between sets of variables present in the data set [48]. This helps in judging which parameters are affecting the output and are they eventually required for building a model. Here, the Pearson's coefficient is used to determine the correlation between input and output. Pearson coefficient can be defined as the strength of linear association between any two variables. It is a technique of investigating the relationship between the variables. The general formula used to calculate the correlation between x and y variables using Pearson's coefficient can be described as Eq (3)

$$\frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}} \quad \text{…….. Eq (3)}$$

Relative Importance method

Johnson [49] explains that the relative importance is the contribution of each variable to the prediction criteria by itself and the combination of other predictive variables. To calculate the relative importance, there is a requirement of partitioning of variance(R squared) among other predictors which includes all variables in data set. However, if there are completely uncorrelated predictors, the relative importance is calculated by dividing the squared standardised regression coefficients (β^2) from the squared multiple correlation(R squared).

Further, if the predictors are correlated, the squared standardized regression coefficients (β^2) no longer sum to the R- squared [49].

Random forest method

This method can be used efficiently to select the predictors by explaining the variance of each parameter relative to the response variable. Selecting the variables from tree derived importance is pretty straight forward and accurate way to select the good features for machine learning [45]. The way this method works is that the random forest algorithm evaluates subsets of multiple features rather than the individuals. It randomly chooses features by building an algorithm out of simple classifiers hence achieving good number of possible feature subset. As not all the training data is included in the constructions of the base hypothesis, the out of bag data provides with the evaluation of feature subsets without having a need of another independent test set [50].All these methods are implemented to the scaffolding data, and the parameters are further grouped as primary and secondary impacting factors, depending on their scores and the comparison of the values obtained from the different feature selection methods. Based on the variable importance, clustered datasets were created to build the predictive models.

To apply the feature processing techniques in this research, the very first step is to cluster the data. The data is clustered into multiple data sets based on the categorical variables. Each work classification is assigned to a discipline and a scaffold type. Out of work classifications-3, Discipline – 7, Scaffold Type -9, there were 189 subsets (3*7*9) that can be formed. Each of the categorical types which had very less percent(less than 3%) of the manhours were grouped all together as miscellaneous field and it was done for each of the categorical parameter, which resulted in the reduction of subsets. The work classification remained with 3 types, however in the discipline category fields such as civil (2% of man hours) and structural (3% of man hours) were merged as miscellaneous and in the scaffold type category the fields (bridge, barricade

and shelter) were merged as miscellaneous. Altogether, there were work classification -3, discipline-5, and scaffold type-6, there were 90 subsets (3*5*6). The 90 subsets then underwent feature selection process. An example of how the clustering is carried out is visually represented in Figure 28.

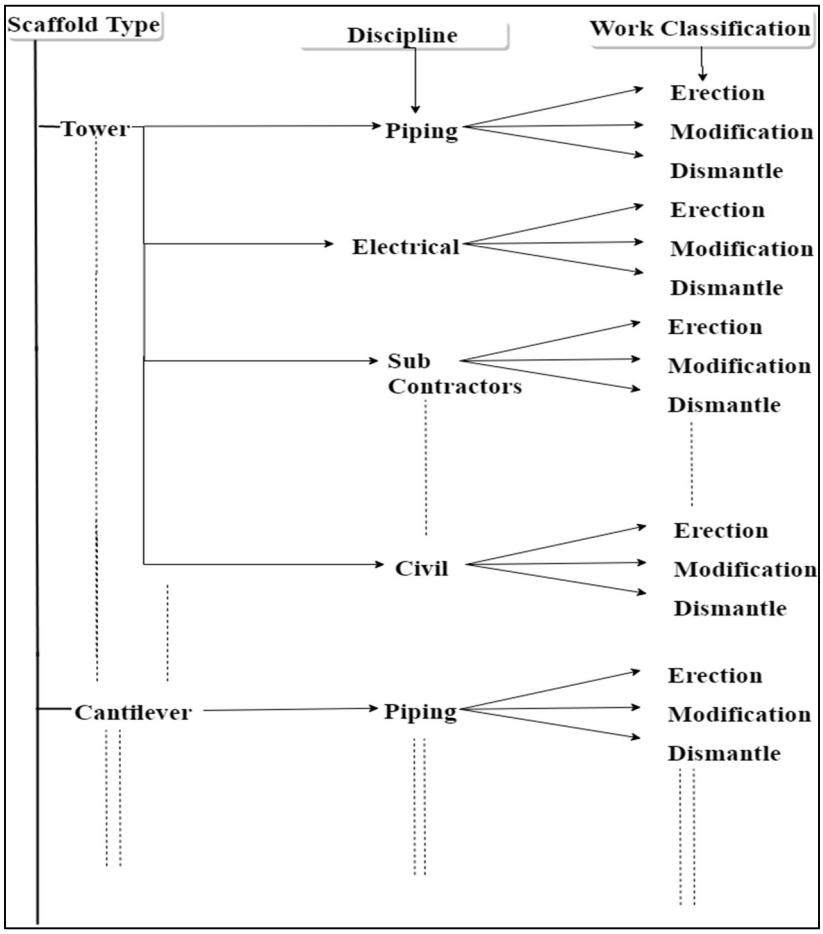


Figure 28 Sample of clustering of data sets

There are three different feature selection process adopted for this research- correlation matrix, relative importance, and the random forest methods. According to the rule of statistics, it is advisory to choose multiple methods to find the importance of variables, so that it can be compared and crosschecked whether the dataset provides the same level of importance for the parameters by using different techniques. It also helps in deciding the parameters for future predictive analysis.The correlation matrices is derived to check which parameters affect the scaffolding manhours (output).Table 4 below represents the correlation between the manhours

and other quantitative parameters present in the data set. The table conveys that the manhours are highly related to the weight of the scaffolding materials, the number of major pieces of scaffold material used, and workable area available for the scaffolding work. There were factors such as temperature, elevation, apprenticeship ratio and night-time ratio which had minor impact on man hours.

Table 4 Correlation for quantitative parameters

	Man hours	Weight of the scaffold	Number of major pieces	Workable Area (Sqft)	Man Count	Apprentice Ratio	Night Time Ratio	Aluminium percent	Temperature	Elevation
Man hours	**100%**									
Weight of the scaffold	**71%**	**100%**								
Number of major pieces	**69%**	**93%**	**100%**							
Workable Area (Sqft)	**69%**	**76%**	**71%**	**100%**						
Man Count	**9%**	**52%**	**53%**	**49%**	**100%**					
Apprentice Ratio	**13%**	**16%**	**8%**	**19%**	**8%**	**100%**				
Night time Ratio	**-7%**	**-11%**	**-4%**	**-15%**	**-27%**	**-68%**	**100%**			
Aluminium percent	**1%**	**8%**	**9%**	**6%**	**29%**	**16%**	**-4%**	**100%**		
Temperature	**4%**	**7%**	**10%**	**6%**	**5%**	**7%**	**16%**	**4%**	**100%**	
Elevation	**5%**	**-2%**	**2%**	**1%**	**5%**	**-8%**	**8%**	**13%**	**8%**	**100%**

The other feature selection processes such as relative importance and random forest method were executed and the relative ranking of the impacting parameters were derived .The test results of the random forest method do indicate that weight, major pieces, workable area has high level relevance to manhours which is tabulated below in Table 5 and the feature selection process done by relative importance method also ranks weight and major pieces has higher relevance represented in Table 6.

Table 5 Random forest method ranking of variables

Parameters related to scaffolding man hours	Percentage of relation
Weight	79.86
Majorpieces	75.11
Workable Area (Square feet)	72.65
Mancount	68.56
Aluminum	62.55
Apprenticeship ratio	-15.71
Elevation	8.85
Night time ratio	-0.20
Temperature	0.48

Table 6 Relative importance method ranking of variables

Parameters related to scaffolding man hours	Importance range
Weight	0.2078
Majorpieces	0.1956
Mancount	0.1780
Workable Area (Square feet)	0.1725
Aluminum	0.0896
Apprenticeship ratio	0.0092
Night time ratio	-0.0059
Elevation	0.0019
Temperature	0.0016

Based on the feature selection processes combined, a list was developed to differentiate the level of importance for each parameters that are affecting the scaffold man hours. There were three division of parameters - primary, secondary and all the parameters impacting the output (scaffold man hours). The primary importance parameters are the ones which show relatively

high importance in all the three feature selection processes. The secondary importance parameters are the ones which repeatedly have higher importance (less than primary) in one or two methods, and at last all those parameters which even have the slightest impact on manhours are considered for the trial and error methods. Each set of parameters are trained separately and combined. There are multiple trial and error combination of parameters executed based on their importance levels. The Table 7 below represents the importance of each parameters in respect to the scaffolding man hours.

Table 7 Importance level of different parameters on scaffolding man hours (Based on feature selection)

Level of Importance	Parameters
Primary Importance (Parameters which are in top 5 scores in all the feature selection methods)	Weight Majorpieces Workable Area Square feet
Secondary Importance (Parameters which were repeated in two or more feature selection methods)	Elevation Man count Aluminium percent
All parameters affecting man hours (Complete list of parameters which have slightest impact on man hours)	Weight Majorpieces Workable Area Square feet Elevation Temperature Man count Apprenticeship ratio Night time ratio

The different data subsets generated is further tested with the regression models and the neural networks until the desired accuracy is obtained .The next chapter completely explains about the individual results of each algorithm in terms of evaluation metrics and accuracy values.

3.6. Modelling

After visualising the factors of the data by the help of bar graphs, pie charts, project life cycle analysis and correlations; the next step is to process the data and build models using machine learning algorithms. Machine learning is an integral part of data science. It is an application of the artificial intelligence where the system automatically learns patterns and improves its performance without programmed explicitly by humans. The process of learning begins from providing observational data and making better predictions from it. The main advantage of using machine learning is that, it makes use of the all the data available in order to provide with efficient results benefited for future uses. In addition, machine learning algorithms also determine the amount of data that can be minimized ,which in turn helps in providing a global solution for optimization of the existing data [51].There are different types of machine learning algorithms used to build the models. Several factors such as data set size, business requirements, and quality of the data, training time, data points and more are considered while choosing a machine learning algorithm. It can be added that selecting a right algorithm is a combination of both the business requirements and the data specifications [52].There are three main divisions of algorithms-*Supervised learning, Unsupervised learning, Reinforcement learning*. Each of them are explained in brief.

i. ***Supervised Learning -*** These algorithms are the most easy to understand and commonly used algorithms. As the name suggests, it makes use of labelled datasets which acts as supervisor during the training. In other words, a dataset in this algorithm is treated as a set of sample and each set of sample is associated with an expected output /label [52].Examples are linear regression, random forest, support vector machine algorithms.

ii. ***Unsupervised learning*** – These are the algorithms built to deal with unlabelled data .The machine has to learn the data by itself and find the intrinsic patterns. This include clustering data of similar patterns, reducing the dimension of data set by removing certain irrelevant

features that does not contribute to the task to be performed[52].Examples are k-means clustering, apriori algorithm for association rule.

iii. ***Reinforcement learning*** – These algorithms works in a way where they analyse and optimise the behaviour of the data based on the feedback provided by the environment. The machine itself interprets different scenarios to figure out which action yields the best outcome rather than being told to the machine in prior. This type of learning is slow since there has to be a lot of trial and error done, but they yield good results [52].Examples are neural networks ,deep deterministic policy gradient.

Generally, in supervised learning; there are three types of models built; (i) *Predictive models*-Analysing from the past information for future predictions, (ii) *Descriptive models* – Classifying the relationship of different parameters into different datasets, and (iii) *Decision models*-To predict the result by the decisions taken [53]. Current study focuses on predictive models built using R Studio software. There are various programming languages to perform data analysis such as C++, Java, Python, R programming etc. In this research, R language is used because it has a lot of inbuilt libraries which helps in building models quickly and efficiently. The initial step in building a machine-learning model is to define the input parameters and the targeted output parameter. To execute any machine learning algorithm, the targeted value (output) in the dataset should be dependent on all the other parameters (inputs) in the data set. Further, the data set is divided into training set (80% of data preferred) and testing set (20% of the data). It is a good practice to train multiple algorithms to figure out which one provides the best fit. The accuracy of each model is tested by introducing series of new data sets and testing the output obtained with the actual values. Linear regression is the simplest types of predictive model where in a linear relation between two variables is established and checked how they are dependent on each other. Further, for complex modelling where in huge amount of data exists, extended regression algorithms such as random forest

regression, decision tree regression and neural network models is used. Each algorithm used in this research is explained.

3.6.1. Multiple linear regression

Multiple linear regression is the extension of simple linear regression, which attempts to identify the value of one dependent variable which is defined by two or more independent (predictor) variables [54]. The regression analysis model built is defined as Eq (4)

$$Yi = B_0 + B_1x_{i1} + B_2x_{i2} + ... + B_px_{ip} + €, \text{ where } i = 1, 2... N. \quad \text{.....Eq (4)}$$

Y_i is the predicted outcome(dependent variable) which is mathematically related to independent variables (x_{i1}, x_{i2}, x_{i3}, x_{i4}) with different regression coefficients (B_0, B_1, B_2, B_p) and € is the residuals [54]. A fitted regression line is generated by applying least square method which is considered as the most common estimator or predictor in regression analysis. The least square method predicts the unknown constant in the hypothesized equation which helps in minimizing the sum of squared deviations of the fitted values from the actual values, therefore determining the accuracy of the model [55].

3.6.2. Decision tree regression

As the name depicts, this model provides with regression or classification models in the form of a tree structure. The way it works is that, while developing the decision tree at every step, it also tends to break down datasets into smaller subsets every time [56]. Non-linear interactions between the input and the output can be captured using this regression model as the decision tree regression model is better at managing and handling tabular data with numerical features, or categorical features [57].

3.6.3. Random forest regression

Decision trees are stand-alone models whereas random forests are an ensemble of multiple decision trees. Random forest regression models are additive models which predicts based on combining the decisions of different sequence of the base models generated. Eq (5)

$$G(x) = f_0(x) + f_1(x) + f_2(x) \ldots + f_n(x) \quad \text{.....Eq (5)}$$

Where G(x) represents the final regression model and series f_0, f_1, f_2...f_n represents the simple base models obtained by running decision tree. Model ensemble is done through the inbuilt programming software which merges the two algorithms. Base models are constructed independently by sub samples of the data generated automatically [58].

3.6.4. Artificial Neural Network (ANN) regression

K. Gurney [59], a profound data scientist defines neural network as

"An interconnected assembly of simple processing elements, units or nodes, whose functionality is loosely based on animal neuron. The processing ability of the network is stored on interunit connections strength s or weights, obtained by the process of adaptation to, or learning from a set of training patterns."

In the world of growing machine learning algorithms, the artificial neural network is another such technique that adopts the work structure of a neuron in a human body. The neurons work based on the external stimuli (input) received and then by performing its activities (function) passes an end signal (output). The neural network is inspired by the functionality of neuron cells. Artificial neural networks is a machine learning algorithm, in which the system learns to perform some task by analyzing the information or the data received and performs a mathematical function to give desired output[60]. The examples of artificial neural network includes image recognition, where the system finds images through visualization and check whether it correlates with particular labels generated. ANN is a set of connected neurons

organized in three layers such as input layer, hidden layer and output layer. Input layer is the one where the initial data fed to the system to process the subsequent layers of artificial neurons. The hidden layer is the layer between input and output layers where artificial neurons take weighted inputs and activation function is set up. Then, the output layer is the one which provides the desired output. Unlike linear regression, neural networks don't assume any relationship (linear /nonlinear) between input and output parameters. Every layer has a random amount of nodes associated with it. For example, the number of nodes in the input layer is equivalent to the dimension of the input data features. When it comes for the number of nodes in the output layer, it depends if it is a regression problem (where the number of nodes is 1) or a classification problem (where the number of nodes is equal to a number of classes or groups that are always greater than 1). Certain matrix operations that include multiplying input data with weights associated in the network layer are executed, post that bias is added to each layer at every node. Afterward, the activation functions such as sigmoid, tan hyperbolic, linear functions is then added to those layers before transferring them further. The activation function is a mathematical formula applied to the inputs along with the weights to provide desired outputs [59]. A bias is an additional node in the hidden layer used to enhance the output and give a better fit. An example of the artificial neural network built is represented in Figure 29.

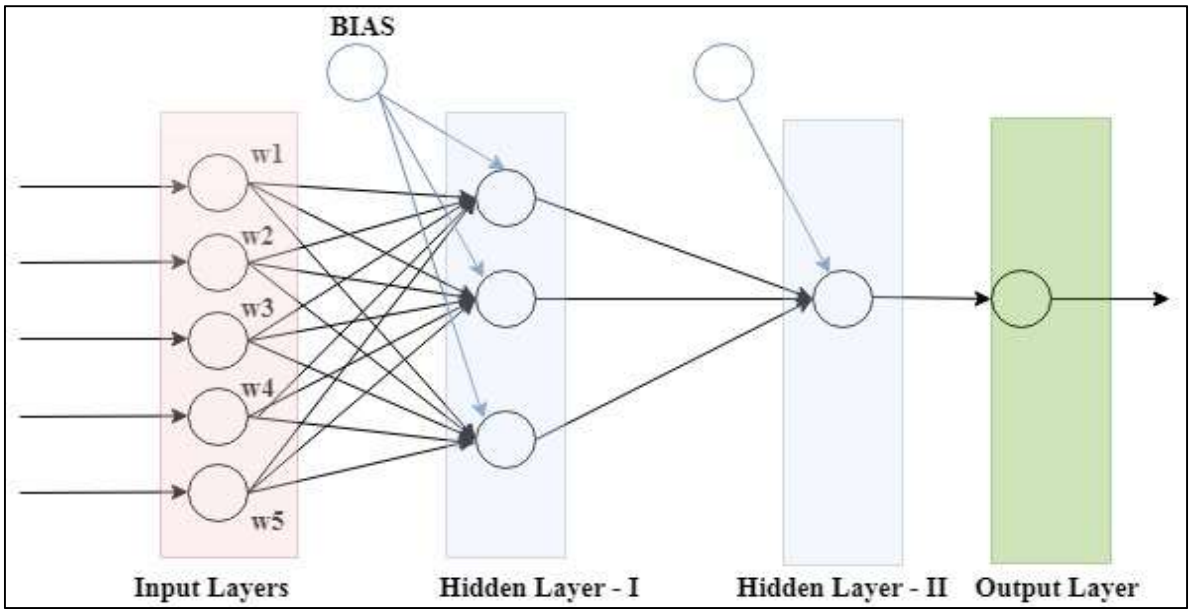


Figure 29 Example of an artificial neural network

While training the artificial neural networks it is advised to maintain a normal distribution for the values of the input parameters to avoid data redundancy. If the values of each parameters in the data sets are in different ranges, certain algorithms does not perform well. The machine learning algorithms perform well when there is uniformity in the range of data. The rescaling of data is carried out to get uniform distribution. The rescaling can be done by two methods- data standardisation and data normalisation. Data standardisation is a rescaling technique which converts the entire data into Gaussian distribution having a mean of 0 and standard deviation ranging 1[61].The general formula for data standardisation is Eq (6)

$$Z_i = \frac{X_i - mean(x)}{std(x)} \qquad \text{..... Eq (6)}$$

z_i corresponds to the ith standardised value and x_i represents all values.

Data Normalisation is to scale the values of all the parameters of different range in the dataset to the range of minimum 0 and maximum of 1[61]. The general formula for data normalisation is Eq (7)

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad \text{.....Eq (7)}$$

z_i corresponds to the *i*th normalised value and x represents all values.

3.7. Model Evaluation

The predictive model built is evaluated by different testing methods to confirm whether the model works well for different data sets, is there any overfitting (explained in Chapter 4) of the model, and to check the error difference in different algorithms. The very first step is to run the model for new test data and check the model's accuracy .There are many statistical methods used to evaluate the models. The most preferred evaluation metrics for regression type analysis are root mean squared error (RMSE), mean absolute error (MAE) and R squared .In this case study , these evaluation metrics are the deciding factors for choosing the best fit model.

3.7.1 Root Mean Squared Error (RMSE)

RMSE is the root square of the average squared distance between the actual and predicted observations. It is one of the crucial metric that is used on the models, especially when it is a regression model. Since the errors are squared, the metric provides high weight to larger residuals (different between the actual and predicted value) than the smaller residuals [62]. Eq (8) represents the general RMSE formula.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2} \quad \text{.....Eq (8)}$$

3.7.2 Mean Absolute Error (MAE)

MAE provides with the average number of errors obtained from a group of estimates without emphasizing on its direction. In other words, it provides with an average of absolute differences observed between the actual and predicted values having equal weight for each difference [62]. Eq (9) represents the general MAE formula.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \quad \text{……..Eq (9)}$$

3.7.3 R squared value

R-square provides with a computation value of how close the data is plotted to the regression line. The R squared is also refereed as the coefficient of determination. If we consider x & y to be independent and dependent variables, then the coefficient of determination provides with percentage variation in y defined by all the x variables together. The results of R square is measured between 0-100%, with 0% meaning that the data in consideration is entirely inconsistent and 100% meaning that the data considered is entirely consistent making it as the perfect model. R squared equation is expressed as (Eq 10)

$$R^2 = SSE/SST \quad \text{……..Eq (10)}$$

Where, SSE = Sum squared regression error = $(y - \hat{y}_j)^2$

SST = Sum squared total error = $(y - \overline{y})^2$

$\hat{y}_j$ = Predicted values, $y =$ Actual values $\overline{y}$= mean of actual values.

Although both (RMSE and MAE) error metrics hold the same point and play a vital role, MAE is considered less popular because instead of pointing out the differences between the performances of models in different situations (unlike RMSE) it makes them appear to the same. MAE consider absolute difference of observed and actual value. This might not be appropriate for all mathematical models. As the R-square provides with a statistical measure of the closeness of the data to the fitted regression line, it can be used for linear and random forest regression approaches. So, higher the R square better the model [63]. Further, all these evaluation metrics are used to judge the predictive model built for scaffold man hours. In the next chapter, implementation of the methodology studied so far, is discussed.

3.8. Implementing the model and test results.

To summarise, the scaffolding data received from the case study of the heavy industrial project is carefully studied to perform preliminary data analysis. Then, through data visualisation, more insights to the available data is extracted. Further, the factors affecting the man hours consumed for different scaffolding activities are represented. After all the preprocessing, the data is set into right format as required for machine learning. To build the predictive models for scaffolding manhours, the processed data is trained by multiple machine learning algorithms and tested for future predictions. Also, as discussed earlier the data is divided into two sets, namely training set and test set. Generally, the split will be 80 percent of the data for training set and 20 percent of the data for test set. The division or split is carried based on number of manhours in this case, since the man hours is the desired outcome. After the training set data is trained by different algorithms, it has to be tested by the new inputs. Hence, the test data is applied to the trained models built by different algorithms to crosscheck their performance. In this research, the training and building process of the model is executed with the help of R studio and R programming. The R programming language has multiple in built codes and libraries such as previously random forest, neural nets, nnets, and caret used for building the algorithms. These libraries help in training the data in a quick and easy manner. There is a set of parameters initially considered to perform the analysis (Table 8). The parameters are free from outliers and the project centric information or data. Before applying the algorithms to the data, the data has to be in its processed state. The processed or the cleaned form of the data includes eliminating the parameters not affecting the output, clearing all the unwanted or empty data points (outlier's removal) and having uniform distribution of parameters.

Table 8 Parameters considered for modelling in the initial stage

Parameters considered Initial stage.		Data types	Range/ Types	Comments
Input	Man Count	numerical	1 -117 members	Man-count helps in deciding the man-hours which help in understanding its requirement for longer projects.
	Temperature	numerical	(-34.1) to 22.9 degrees	Productivity varies based on temperature especially in extreme temperatures (Cold or Hot)
	Elevation	numerical	1-110 ft.	Higher the elevation lesser the productivity since working in heights consume time.
	Apprenticeship Ratio	numerical	0-64%	Work done by a skilled worked against an apprentice (trainee) is quite different.
	Night-time Ratio	numerical	0-78%	Productivity during the day is far more than during the night.
	Total Weights	numerical	24-125640 lbs	Weight of the materials directly influences the labor.
	Major Pieces Count	numerical	1-6732	Major pieces count uses more labor than minor pieces, making it an important contributing factor.
	Workable Area (Sqft.)	numerical	0-11063 sq. ft.	Productivity is directly impacted with the amount of space available for the labor to work.
	Work Classification	categorical	3 types	Erection, Modification and Dismantle.
	Discipline (Trades)	categorical	7 types	Piping, Electrical, Civil, Subcontractors, Mechanical, Structural and General Scaffold.
	Scaffold Type	categorical	8 types	Tower, Barricade, Bridge, Cantilever, Working Deck, Dance floor, Hanger and Hoardings.
Output	Manhours	numerical	1-2000 hours	

The variations in the categorical data present is also considered as the disturbance to the data pattern, since some trades had larger affects (more man-hours) compared to the other. The unevenly distributed categorical parameters (disciplines, work type and other categorical data) might be the reason for the poor performance of data. Hence, it was decided to try sub setting of the data or clustering of data into groups based on different disciplines, work types and scaffold types .Splitting of data made it more specific. The proposed models was tried in 5 different approach.

- Complete data set (Considering the dummy variables for categorical data
- Grouped data set based on Work Classification.
- Grouped data set based on Discipline.
- Grouped data set based on Type of scaffold.
- Grouped data set based on Work Classification and Discipline (all numerical).

Each of these divided subsets are trained and tested on different algorithms. The results are further discussed in this section.There were multiple iterations tried on different clustered datasets to build the model.The flow chart below (Figure 30) explains how each model is built, validated with the help of new test data and how the best model is chosen. Initially, the data set is split into two sets; training set and test set. As discussed, the training of data points in the dataset is done by different machine learning algorithms. The models built by similar research works from C.Kumar [7] and L.Wu [19], yields result up to 75-80%.Keeping that as a bench mark, the current research, targets to a minimum of 90% of accuracy in order to provide the best fit model. Also, the model can be used for real time data in future for which maintaining high level of accuracy is very important. Hence, each of the algorithms are implemented in multiple ways and results are discussed to choose the best model.

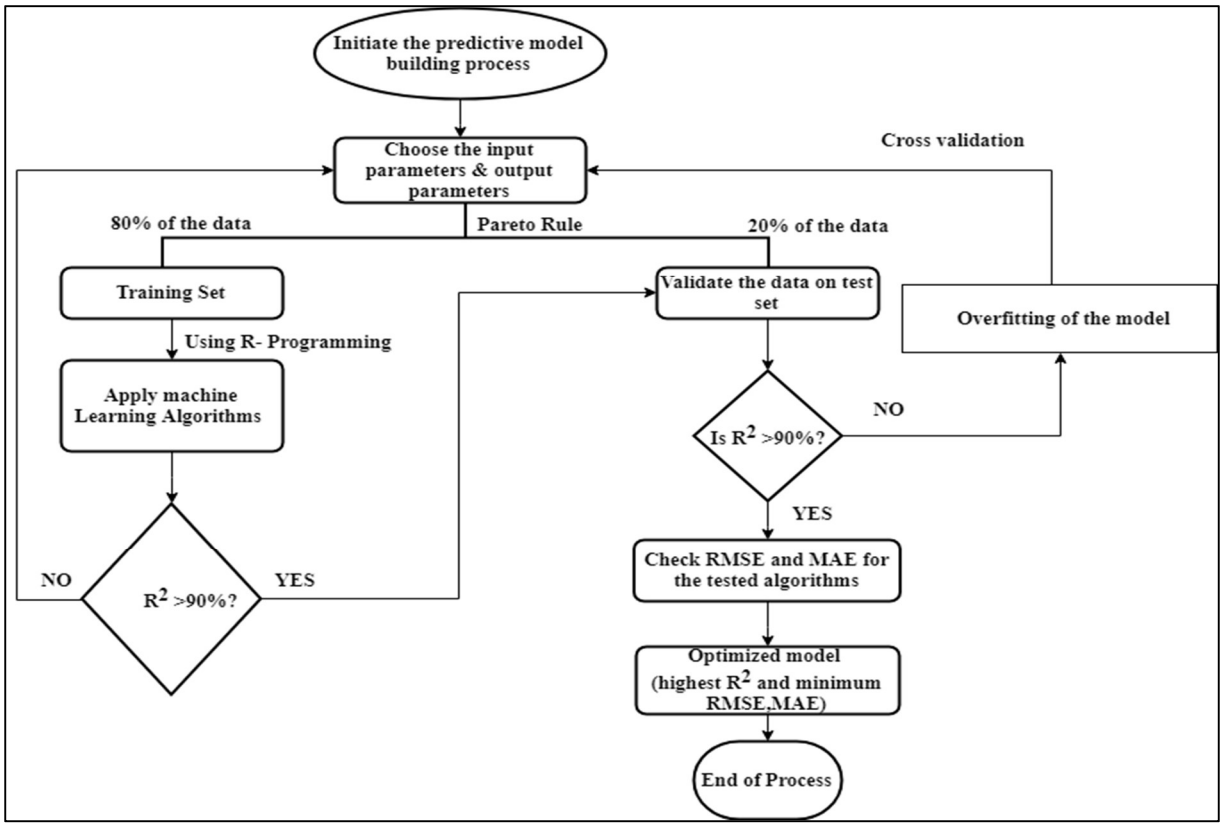


Figure 30 Flow chart of building a predictive model.

There are certain times wherein few algorithms work well for training data, but then they fail to give good results for new data or test data. This situation where the model is unable to perform well on the newer data as against its performance on the training data is called overfitting [64]. A good model will learn the pattern from the training data and then generalize it on new data (from a similar distribution). Random forest or neural net can easily over fit. To detect overfitting, we need to see how the test errors evolve. If the test error is decreasing with new set of inputs, the model is still right. On the other hand, an increase in the test error indicates that the algorithm is probably overfitting. In such situations, it's always recommended to reconsider the data set, which means sub setting of data in different approaches.There is no solution other than doing trial and error. Linear regression with a reasonable number of the variable will never over fit the data, since the model is simple and restricted to linear relationships between the variables [65]. After the data is clustered and is in its processed state, the work flow of building predictive model is defined, and the final stage is to apply different

machine learning algorithms and evaluate their accuracy as well as error percentage. The machine learning algorithms used in this research are multiple linear regression, decision tree, random forest regression and artificial neural networks (ANN).The results after applying these algorithms to the clustered data sets of scaffolding data is discussed further. Initially, based on the IQR different ranges defined in the data pre-processing stage, were trained with the machine learning algorithms (Table 9) .The results did not yield the required accuracy of 90 %. Hence, the data set based on company's rule is considered for all the further analysis.

Table 9 IQR different range model analysis

IQR probability range	Number of rows removed from each IQR range	Cumulative Accuracy of different models (%)			
		Multiple linear regression	Decision tree	Random forest regression	Artificial neural networks.
25%-75%	3172	59.03	54.07	60.3	61.8
20%- 80%	2343	64.3	58.6	65.39	68.41
15% - 85%	2052	66.5	61.1	68.2	69.8
10%- 90%	1534	74.8	64.5	72.5	73.5
5% - 95%	388	88.4	67.7	78.5	89.8

Based on the data pre-processing and feature selection techniques discussed, there were attempts made to build model for scaffolding manhours using datasets with only parameters of primary level of importance. However, the test results from less number of input parameters did not have significant measure of accuracy (ranged around 55-60%) and the error values turned out to be very high. Further, secondary level of importance were added to the datasets to check if there is an improvement in the accuracy and decrease in the error percent. Unfortunately, there were no significant changes. The test results appeared almost same as they appeared for datasets which were tested with primary level of importance. Further, the backward propagation method was adopted to build the predictive model using all the parameters which were related to output variable in both higher level and lower level which

means the ones with minimum correlation were also considered to build model. The test results increased and almost reached the target of 90% accuracy (Table 10).

Table 10 Predictive models based on variable importance

Algorithms		**Primary Importance**	**Secondary Importance**	**All Factors**
		Weight Majorpieces Workable Area (Sqft)	Weight Majorpieces Workable Area (Sqft) Elevation Temperature Man count	Weight Majorpieces Workable Area (Sqft) Elevation Temperature Man count Apprenticeship ratio Night time ratio
Multiple linear regression	RMSE	78.36	68.36	48.06
	MAE	33.6	30.2	28.15
	R^2 (%)	65	75.6	88.4
Decision tree regression	RMSE	123.57	102.6	75.6
	MAE	78.5	68.3	46.56
	R^2 (%)	58	59.9	65.6
Random forest regression	RMSE	98.9	88.3	68.65
	MAE	67.6	56.89	37.9
	R^2 (%)	69.56	70.56	76.46
Artificial neural networks	RMSE	100.56	85.8	25.86
	MAE	38.6	34.67	16.85
	R^2 (%)	77	79.2	91.2

As discussed earlier, the proposed model was tried to be built by 5 different divisions of the data based on the work classification, discipline and scaffold type. The test results of each division is tabulated from Table 11 to Table 14.

Table 11 Test results of complete data set

Data Set	Algorithms	Test Results		
		RMSE	MAE	R squared %
Complete Data Set	Multiple Linear Regression	105.75	63.58	62.3
	Random Forest	111.91	47.44	64.73
	Decision Tree Regression	129.17	63.68	58.05
	Neural Network	119.21	49.91	62.59

The accuracy level for complete dataset ranged around 50-60%. The next proposed approach was grouping based on type of scaffold (Table 12). The accuracy of this clustered data set also ranged in between 50-60% which was same as grouping done for the complete data (Table 11). The reason for the similar result was, in the type of scaffold only tower scaffolding has 69% of the data. Therefore, while grouping based on scaffold type, most of the data points would get dominated in one single sub set (Tower scaffold). When one particular scaffold type is occupying almost 70% of data, then there is no point in taking it as a decision criteria since it neutralises other types of scaffold. Hence, type of scaffold was ignored in this case study.Further, the training of clustered data based on Work Classification (Table 13) and based on discipline (Table 14) was done. It was observed that there was a rise in the accuracy level (65-85%).Also, random forests worked well compared to any other machine learning algorithms.

Table 12 Test results for dataset grouped based on type of scaffold

Data Set	Algorithms	Test Results		
		RMSE	MAE	R squared %
Data Set grouped based on Type of Scaffold	Multiple Linear Regression	94.86	58.76	66.23
	Random Forest	109.4	43.54	65.8
	Decision Tree Regression	148.95	75.56	53.78
	Neural Network	120.81	56.8	63.48

Table 13 Test results for dataset grouped based on work classification

Data Set	Algorithms	Test Results		
		RMSE	MAE	R squared %
Data Set grouped based on Work Classification	Multiple Linear Regression	69.3	43.8	83.6
	Random Forest	107.93	50.01	83.85
	Decision Tree Regression	145.71	78.85	68.06
	Neural Network	114.53	53.60	73.6

Table 14 Test results for dataset grouped based on discipline

Data Set	Algorithms	Test Results		
		RMSE	MAE	R squared %
Data Set grouped based on Discipline	Multiple Linear Regression	68.4	47.25	83.5
	Random Forest	101.65	47.44	84.56
	Decision Tree Regression	129.17	73.68	66.8
	Neural Network	137.21	59.25	72.9

The final grouping of parameters were completely numerical in which multiple subsets were formed (as in for each discipline and each work classification there was a separate data set) .The completely clustered group yielded better results compared to different grouping done so far. The accuracy was tested based on RMSE, MAE and R squared value obtained. The parameters, which worked well for regression models, did not give appreciable results for neural networks. Further, number of iterations were tried for neural networks by changing the hidden layers, changing the number of neurons in each layer and removing certain parameters, but the performance of the neural network didn't improve. After trying multiple trial and error data sets, a new approach was adopted in which more weight was added for the parameters which are highly related to the output. This approach finally yielded good results for ANN algorithm. The final set of parameters which were selected for the best-fit models are mentioned in the Table 15.

Table 15 Final set of input parameters for the predictive models

Dataset for regression models	Dataset for ANN models
Elevation	Elevation
Temperature	Temperature
Man count	Man count
Workable area(Sqft)	Workable area
Weights of scaffold	Weights of scaffold
Night time ratio	Night time ratio
Apprenticeship ratio	Apprenticeship ratio
Number of major pieces	Number of major pieces
	Weights of scaffold ^2
	Number of major pieces^2
	Workable area(Sqft)^2

In Table 16, we can notice that the results of the regression models are quite in the acceptable range. The multiple linear regression ranges around 85-90% of accuracy, random forest with 80-85% and decision trees about 60-70%.Even after trying 4-5 iterations the results clearly states that the multiple linear regression is giving more accurate results. Though there is a

gradual increase in random forest regression model from 79- 87% it failed to reach the target. Similarly decision tree regression model, which performed poorly from the start could reach maximum of 75%. It is said that a model always performs well by training it in different patterns and doing the necessary trial and error methods which are statistically approvable. The whole process of data analysis will consume time to provide better outcomes.

Table 16 Cumulative test results for regression models

Regression Models									
	Multiple Linear			***Random Forest***			***Decision Tree***		
Iterations	***MAE***	***RMSE***	R^2	***MAE***	***RMSE***	R^2	***MAE***	***RMSE***	R^2
1	73.2	103.5	0.82	63.45	150.6	0.79	79.56	173.68	0.65
2	36.89	82.34	0.85	54.13	123.09	0.82	51.03	160.54	0.65
3	30.02	54.35	0.88	52.89	105.68	0.82	58.5	116.25	0.67
4	25.48	45.08	0.901	41.02	73.14	0.87	62.45	103.8	0.75

However, the artificial neural networks perform differently. As discussed earlier, the input parameters used for artificial networks were different compared to regression models. The number of neurons for each layer must be provided one by one. There is no fixed rule again for providing the neurons .After trial and error by adding hidden layers from 1 to 10 and then changing the number of neurons (n= 1 to 10) for each layer, only one particular set gave results crossing the target accuracy level of 90%. By analysing the results of neural networks, it is noticeable that they have the least error when compared to the other regression models. Table 17 represents that the neural network with two hidden layers with neurons = 3 and 1 gives the good result i.e., high accuracy and less error. All the additional layers were just to check if the accuracy level will increase based on adding and removing neurons for every single hidden layer.

Table 17 Test results of neural networks

Artificial neural networks				
Hidden layers	***n=Number of neurons for each hidden layer***	***MAE***	***RMSE***	***R^2***
2	n=3,1	15.6	24.9	0.91
3	n=4,3,2	28.06	56.8	0.75
4	n= 5,4,3,2	44.02	94.26	0.72
5	n=5,4,3,2,1	56.05	110.36	0.72

Each of these models were tested with the new data for validation. The graph below (Figure 31) shows the collinearity of each model with the actual values. It can be noticed that the values obtained from neural network almost collide with the actual values line, which means the predicted values are almost same as the actual ones. It can be clearly said that artificial neural networks (ANN) are more collinear compared to the other algorithms whose predicted values are dispersed out from the range.

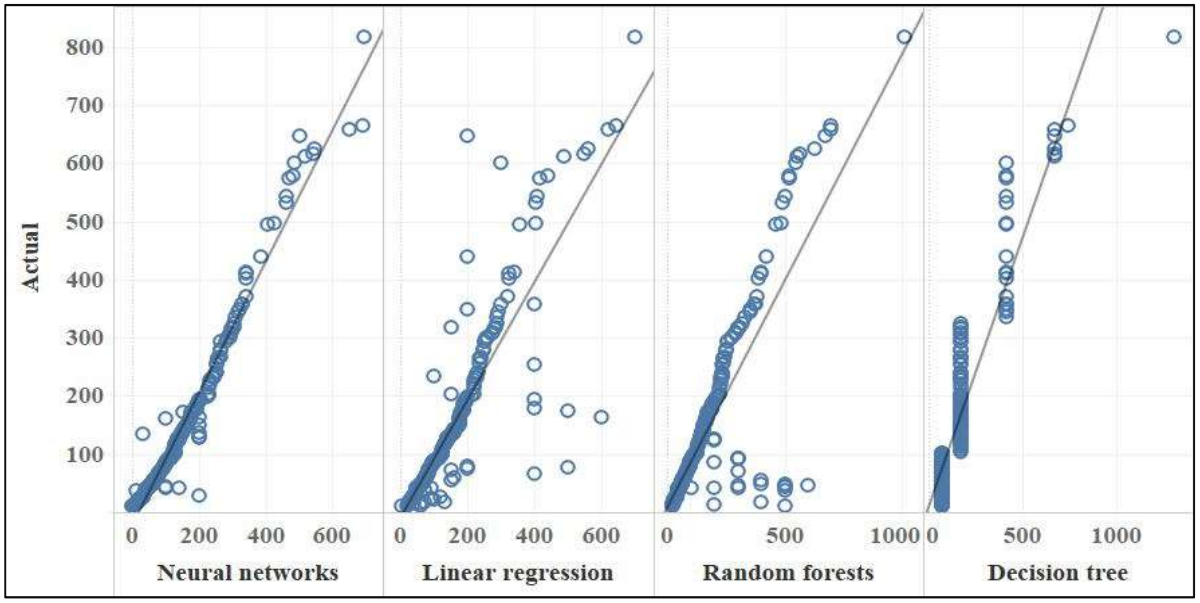


Figure 31 Test results of algorithms in comparison with actual values

Also, by the help of company's officials, an attempt to collect the manhours manually calculated before assigning the scaffolding tasks for the case study. The actual spent manhours, manually estimated manhours and predictive model based manhours were compared. It was

observed that the predictive model built using machine learning algorithm provided better collinearity with the actual values when compared to the manually estimated manhours. This clearly indicates that the computer based systematic approach is a far better option rather than using the decisions of scaffold practitioners as it fluctuates from person to person based on their experience level. Also, the estimation vary from company to company based on the criteria they adopt. The Figure 32 below shows how the machine based prediction of manhours using neural networks (because it has the least error) is far better than the manually estimated values.

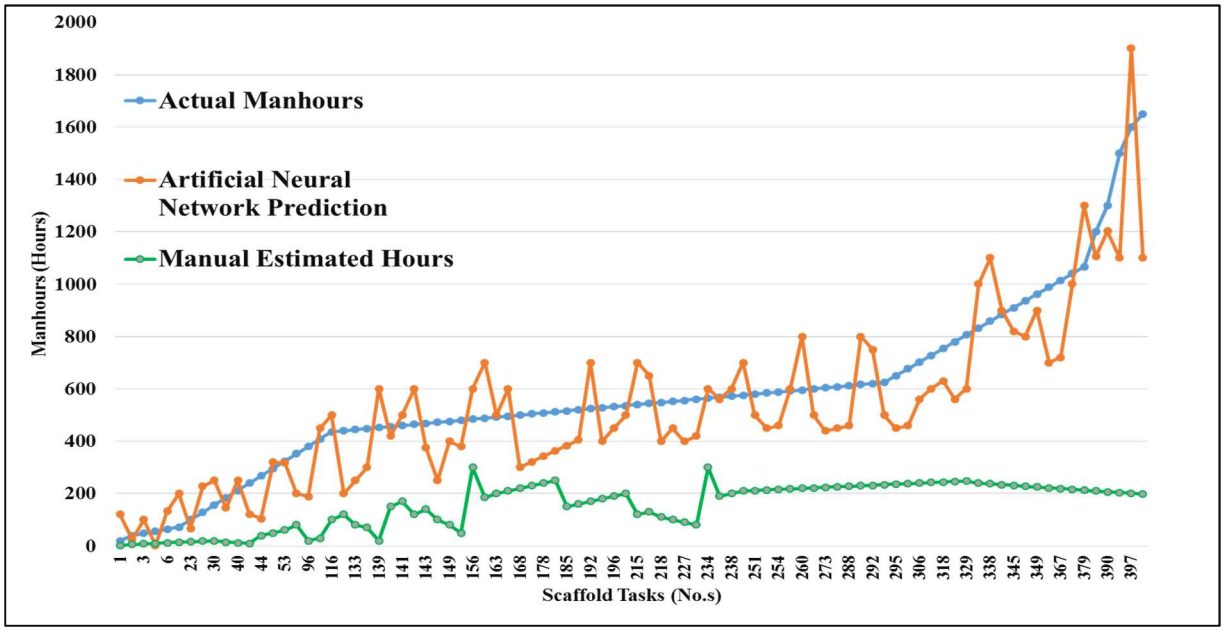


Figure 32 Test results of artificial neural network predictive model in comparison with manual estimation and actual values.

CHAPTER 4: CONCLUSIONS

This research briefs about the importance of scaffolding planning in the starting phase of the project especially in heavy industrial projects which invest more capital and resources even for the temporary works. Scaffolding is being used for all the trades and it is consuming significant amount of labor and materials. There is a need of proper scaffolding planning to avoid unnecessary schedule delays and cost over runs. Any problem in the scaffolding works would affect the direct works of any project considered, because they are the main supports for permanent structures. The current situation in the construction domain has scarcity of labor resources. This motivates for effective utilisation of the man power. The research aims to figure out which kind of tasks consume more man hours for scaffolding, what trade is associated with it, which type of scaffold is consuming more time. Also, the level at which productivity of a scaffolding activity varies with different factors such as temperature, elevation, man count and other parameters is briefly analysed in the first stage of this research work. Careful observations have been made for each parameter associated with the scaffolding man hours and productivity. The main goal of the research was to find out a scientific and a systematic method of predicting the man hours for scaffolding activities in the future projects, so that there is no room for ad hoc decisions while planning scaffolding. This objective was achieved with the help of machine learning algorithms. Before building a future predictive model, a series of statistical analysis such as consolidating and merging the scaffolding data from multiple resources, removing the potential outliers, adopting feature selection process for choosing the parameters affecting scaffold man hours, data visualisation and many other approaches were carried out to understand the data clearly and convert the data into the required format of data mining. The visualisation of the scaffolding activities workflow along time period, with respect to different kind of trades, work type etc. helped the collaborated company to know where their resources have been used significantly.

Data analysis has been a trending topic in the recent researches of construction sector, because the statistical analysis provides insights of the construction activities that can be used for future purposes. An approach has been made to detect the scaffolding man hours in prior believing that planning is a quintessential requirement for the effective project management. An approximation of man hours required for the scaffolding tasks would provide the supervisors with a better planning and management of the projects in a detailed way. To achieve this, machine learning has been a useful tool to build the scaffold man hour predictive models. The programming language R was used to train the models by machine learning. Algorithms such as Multiple Linear Regression, Decision Tree Regression, Random Forest Regression and Artificial Neural Networks was used in this research. To determine the output (scaffold manhours) by using various inputs (elevation, temperature, weight of scaffold, crew members availability etc.) these algorithms were trained and multiple iterations were done to get a best fit model. The results convey that the multiple linear regression and artificial neural networks performed well compared to other two algorithms (decision trees and random forest regression). On an average, multiple linear regression reached upto 90%. The decision tree regression and random forest regression performance ranged around 75% and 87% respectively. Their performance was below average due to the less accuracy and the high error. However, the end results do clearly say that artificial neural networks hold well with the least possible error and an average of 91% accuracy in the model, which was constructed with 2 hidden layers each having 3 and 2 neuron respectively. The reason why artificial neural networks are chosen as better model compared to linear regression is the RMSE and MAE values for the neural network model is very less. Since, the model has its practical use, multiple efforts were made to get more accurate and less errored model, hence target of 90% was set to achieve a better performing model. The models created are definitely useful for future planning

of scaffolding works in any heavy industrial projects. It is one of the initiative steps, to use machine learning algorithms and data science, in the temporary works of construction field.

4.1. Future Work

Machine learning and data science is a vast topic. There is no limit for trying and approaching different methods. There might be multiple ways to explore the data and give machine optimized models .Few of the future scopes or approaches for the future research in order to plan scaffolding in the planning phase of any heavy industrial projects include

i. Many different machine learning algorithms can be tried in building the machine oriented models for manhours prediction.
ii. The statistical analysis could be carried out in different ways by categorizing the scaffolding data based on different ranges of weights of scaffolds, volume of a scaffold structure or any other parameters.
iii. A comparison of the direct manhours for different trades to the scaffolding manhours can help the estimators to decide and allocate manhours accordingly for each trade.
iv. There are high chances of getting a generalized idea on what parameters the scaffolding man-hours are dependent and what is the standard productivity for each man hour if the productivity analysis is carried out for many other heavy industrial projects undertook by the company.

A key take away is, it is important to mention that tracing all the necessary data points related to any project undertaken will guide and help the company to take the correct decisions. It has been proven from various researchers that, the data driven decisions are more effective than traditional practices which are completely relied on human judgements.

CHAPTER 5: REFERENCES

[1] Build Force Canada. (2019). Canadian Territories – Construction & Maintenance Looking Forward. Funded by the Government of Canada.

[2] Ernst & Young. (2012). Exploring the top 10 opportunities and risks in Canada's oil sands. Ernst and Young LLP. (PP. 1-40).

[3] Sierra Systems. (2011). Unlocking the Prize: Metal Fabrication Procurement for Development and Operation of Alberta's Oil Sands: Alberta Oil Sands Supply Chain Opportunity Analysis. Alberta Finance and Enterprise. (PP. 1-42).

[4] Build Force Canada. (2019). National Summary - Construction & Maintenance Looking Forward. Funded by the Government of Canada.

[5] Ratay, R.T. (2012). Temporary Structures in Construction (3rd Ed.). Scaffolding, Chapter 14, McGraw-Hill Professional. ISBN: 978-0-07-17-5307-4.

[6] Kim, K., & Teizer, T. (2014). Automatic Design and Planning of Scaffolding Systems Using Building Information Modelling. Advanced Engineering Informatics. (Vol. 28.1, pp. 66-80).

[7] Kumar, C. AbouRizk, S. M., & Mohamed, Y. (2013). Estimation and planning methodology for industrial construction scaffolding. M.Sc. thesis, Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Alberta.

[8] Department of Labour. (1994). Occupational Safety and Health Services, Safe Erection and use of Scaffolding. Wellington, New Zealand.

[9] Marks, M. T. (2016). Scaffolding – The Handbook for Estimating and Product Knowledge. ISBN: 978-1-68-348358-8

[10] Equipment – Scaffold, Chapter 21. Infrastructure Health & Safety Association. https://www.ihsa.ca/rtf/health_safety_manual/pdfs/equipment/Scaffolds.pdf

[11] Tubular Steel Frame Scaffolding. Digital image. Saint-pro steel company limited. http://steelscaffold.sell.everychina.com/p-98910926/showimage.html

[12] Detail_of_scaffold_tubes_and_clamps. JPEG file from http://www.constructionphotography.com/ImageThumbs/A052-00750/3/A052-00750_Detail_of_scaffold_tubes_and_clamps.jpg

[13] System-Scaffold. JPEG File. 2015. https://www.aluminiumscaffolds.com.au/wp-content/uploads/2015/12/System-Scaffold.jpg

[14] Scaffolding-tower. JPEG File. 2016. http://trinityhire.com/wp-content/uploads/2016/05/Scaffolding-tower.jpg

[15] Proactive Management Solutions LLC, Scaffolding Services, Pompano Beach, Florida. https://www.proactivems.com/copy-of-services

[16] Ducker & Young, Cantilever Scaffolding works, Witney, Oxfordshire. https://www.duckeryoung.co.uk/cantilever-scaffolds/

[17] Hanging scaffold. *Illustrated Dictionary of Architecture*. (2012, 2002, 1998). https://encyclopedia2.thefreedictionary.com/Hanging+scaffold

[18] Suspended Scaffoldings, Edinburg Scotland, Photo Credit: National Geographic Television. https://scaffmag.com/wp-content/uploads/2010/09/suspended-scaffolds-4.jpg

[19] Wu, L. (2013). Analyzing Scaffolding Needs for Industrial Construction Sites using Historical Data. Education & Research Archive. DOI: https://doi.org/10.7939/R3CQ41

[20] Construction Owner Association of Alberta (COAA). (2013). Advanced Work Packaging. Workface Planning. Document Number: COP-WFP-SPD-16-2013-v1

[21] Hou, L., Wu, C., Wang, X. & Wang, J. (2014). A framework design for optimizing scaffolding erection by applying mathematical models and virtual simulation. Proc., International Conference on Computing in Civil and Building Engineering. (PP 323–330).

[22] Pushkar, A., Senthilvel, M., & Varghese, K. (2018). Automated Progress Monitoring of Masonry Activity using Photogrammetric Point Cloud, 35th ISARC. DOI: https://doi.org/10.22260/ISARC2018/0125

[23] Poh, C.Q.X., Ubeynarayana, C.U., & Goh, Y.M. (2018). Safety Leading Indicators for Construction Sites: A Machine Learning Approach, Automation in Construction. (Vol. 93, pp. 375-386).

[24] Kononenko, I, & Kukar, M. (2007). Introduction, Machine Learning and Data Mining. Woodhead Publishing (Chapter 1, pp. 1–36).

[25] Siqueira, I. (1999). Neural Network Based Cost Estimating. Thesis in Department of Building and Civil Engineering, Concordia University Montreal, Quebec, Canada.

[26] Hammad A., Mohamed, Y., & AbouRizk, S. (2014). Application of KDD techniques to extract useful knowledge from labor resources data in industrial construction projects. Journal of Management Engineering. DOI: 10.1061/(ASCE)ME.1943-5479.0000280

[27] Mohammad, R. H., & Adeli, H. (2018). Novel Machine-Learning Model for Estimating Construction Costs Considering Economic Variables and Indexes. ASCE Library Civil Engineering and Its Practical Applications. DOI: 10.1061/(ASCE)CO.1943-7862.0001570

[28] Wang, W., Chen, J., & Hong, T. (2018). Occupancy prediction through machine learning and data fusion of environmental sensing and Wi-Fi sensing in buildings, Automation in Construction. (Vol. 94, pp. 233-243).

[29] Ghousi. R. (2015). Applying a Decision Support System for Accident Analysis by Using Data Mining Approach: A Case Study on One of the Iranian Manufactures. Journal of Industrial and Systems Engineering, Iranian Institute of Industrial Engineering. (Vol 8, Issue 3, pp. 59-76).

[30] Campbell, K., Clegg, D.R., Perera, T., Stephenson, P., & A. Stevens (1997). Simulation in the Construction Industry-A case study review. 13th Annual ARCOM Conference, King's College, Cambridge. Association of Researchers in Construction Management. (Vol. 2, pp. 408-17).

[31] Feng, C., & Lu, S. (2017). Using BIM to Automate Scaffolding Planning for Risk Analysis at Construction Sites, I.A.A.R.C. - International Association for Automation and Robotics in Construction. DOI: https://doi.org/10.22260/ISARC2017/0085

[32] Hou, L., Zhao, C., Wu, C., & Moon, S. (2016). Discrete Firefly Algorithm for Scaffolding Construction Scheduling. DOI: 10.1061/(ASCE)CP.1943-5487.0000639

[33] Cho, C., Kim, K., Sakhakarmi, S., & J. Park. (2018). Machine Learning for Assessing Real-Time Safety Conditions of Scaffolds. (PP. 56-63). DOI: https://doi.org/10.22260/ISARC2018/0008

[34] Sutrisnal, M., Chai, J., Wu, C., & Wang, X. (2018). Exploring the Potential for Automating Progress Tracking of Scaffolding in Construction Projects. ISBN: 978-1-78321-299-6.

[35] Moon, S., Forlani, J., Wang, X., & Tam, V. (2016). Productivity study of the scaffolding operations in liquefied natural gas plant construction: Ichthys project in Darwin, Northern Territory, Australia. Journal of Professional Issues in Engineering Education and Practice. (Vol. 142(4), pp. 04016008-1-10). DOI: 10.1061/(ASCE)EI.1943-5541.0000287

[36] De Francesco, H. F. (2015). Numbers - See how Math works without working at MATH. ISBN: 978-1-5035-2151-3.

[37] Geng, H. (2017). Internet of Things and Data Analytics Handbook (1st Ed). ISBN: 978-1-119-17364-9.

[38] Garcia, S., Luengo, J., & Herrera, F. (2015). Data Pre-processing in Data Mining. ISBN 978-3-319-10247-4. DOI: https://doi.org/10.1007/978-3-319-10247-4

[39] Han, J., Pei, J., & Kamber, M. (2006). Data Mining - Concepts and Techniques (2nd Ed). ISBN 13: 978-1-55860-901-3.

[40] Negri, R., Carlos, J., & Liboni, L. H. B. (2016). Improving consumption estimation of electrical materials in residential building construction. Automation in Construction. Elsevier B.V. (Vol. 72, pp. 93–101). DOI: https://doi.org/10.1016/j.autcon.2016.08.042

[41] Huang, G., Sun, Y., & Li, P. (2011). Fusion of redundant measurements for enhancing the reliability of total cooling load based chiller-sequencing control. Automation in Construction, Elsevier B.V. (Vol. 20(7), pp. 789–798). DOI: 10.1016/j.autcon.2011.02.001

[42] Qady, M.A., & Kandil, A. (2014). Automatic clustering of construction project Documents based on textual similarity. Automation in Construction. (Vol. 42, pp. 36–49). DOI: 10.1016/j.autcon.2014.02.006

[43] Manikandan, S. (2011). Measures of Dispersion, Journal of Pharmacology & Pharmacotherapeutics. (Vol. 2(4), pp. 315–316). DOI: 10.4103/0976-500X.85931

[44] Iliinsky, N., & Steele, J. (2011). Designing Data Visualizations (1st Ed.). ISBN: 978-1-449-31228-2.

[45] Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection, Journal of Machine Learning Research. (Vol. 3, pp. 1157-118).

[46] Jovic, A., Brkic, K., & Bogunovic, N. (2015). A review of feature selection methods with applications. Opatija, Croatia. DOI: 10.1109/MIPRO.2015.7160458

[47] Hall, M. A. (1999). Correlation-based Feature Selection for Machine Learning, A Thesis in Doctor of Philosophy, The University of Waikato, Hamilton, New Zealand.

[48] Kassambara, A. (2017). Practical Guide to Cluster Analysis in R - Unsupervised Machine Learning (1st Ed.).

[49] Johnson, J. W. (2000). A heuristic method for estimating the relative weight of predictor variables in multiple regression. Multivariate Behavioural Research. (Vol. 35, pp. 1-19). DOI: https://doi.org/10.1207/S15327906MBR3501_1

[50] Saunders, C., Grobelnik, M., Gunn, S., & Taylor, J. S. (Eds.). (2006). Subspace, Latent Structure and Feature Selection. (PP. 173-182). ISBN-10: 3-540-34137-4.

[51] Mueller, J. P., & Massaron, L. (2016). Machine Learning for Dummies. ISBN: 978-1-119-24551-3.

[52] Sosnovshchenko, A. (2018). Machine Learning with Swift: Artificial intelligence for iOS (1st Ed.). ISBN: 978-1-78712-151-5.

[53] Information Resources Management Association. (2017). Decision Management: Concepts, Methodologies, Tools and Applications. ISBN: 978-1-52-251838-9.

[54] Jobson, J. D. (1991). Multiple Linear Regression. In: Applied Multivariate Data Analysis. Springer Texts in Statistics (Chapter 4). Springer, New York, NY.

[55] Golberg, M. A., & Cho, H. A. (2010). Introduction in Regression Analysis, Reprinted Edition. ISBN: 1-85312-624-1

[56] Sayad, S. (2010). An Introduction to Data Science. Computer Science Undergraduate Program Tracks. Department of Computer Science, Rutgers School of Arts and Sciences.

[57] Turi Machine Learning Platform User Guide. (2018). Decision Tree Regression material, Create API 1.10 Documentation.

[58] Turi Machine Learning Platform User Guide. (2018). Random Forest Regression material, Create API 1.10 Documentation.

[59] Gurney, K. (1997). An Introduction to Neural Networks (1st Ed.). Master ISBN: 0-203-45151-1.

[60] Ciaburro, G., & Venkateswaran, B. (2017). Neural Networks with R: Smart models using CNN, RNN, deep learning and artificial Intelligence principles (1st Ed.). ISBN: 978-1-78839-787-2.

[61] Saleh, H. (2018). Machine Learning Fundamentals: Use Python and scikit- learn to get up and running with the hottest developments in machine learning. ISBN: 978-1-78980-355-6.

[62] Rocha, A., Adeli, H., Reis, L. P., & Costanzo, S. (2018). Trends and Advances in Information Systems and Technologies. Advances in Intelligent Systems and Computing. (Vol. 3). DOI: https://doi.org/10.1007/978-3-319-77700-9.

[63] Hodnett, M., & Wiley, J. F. (2018). R Deep Learning Essentials: A step-by-step guide to building deep learning models using Tensor Flow, Keras and MXNet (2nd Ed.). ISBN: 978-1-788-99289-3.

[64] Harrell, F. (2001). Regression Modelling Strategies: With Applications to Linear Models. Logistic Regression and Survival Analysis (1st Ed.).ISBN: 978-1-4419-2918-1.

[65] Pearson, R. K. (2018). Exploratory Data Analysis Using R. ISBN: 978-1-1384-8060-5.

APPENDIX

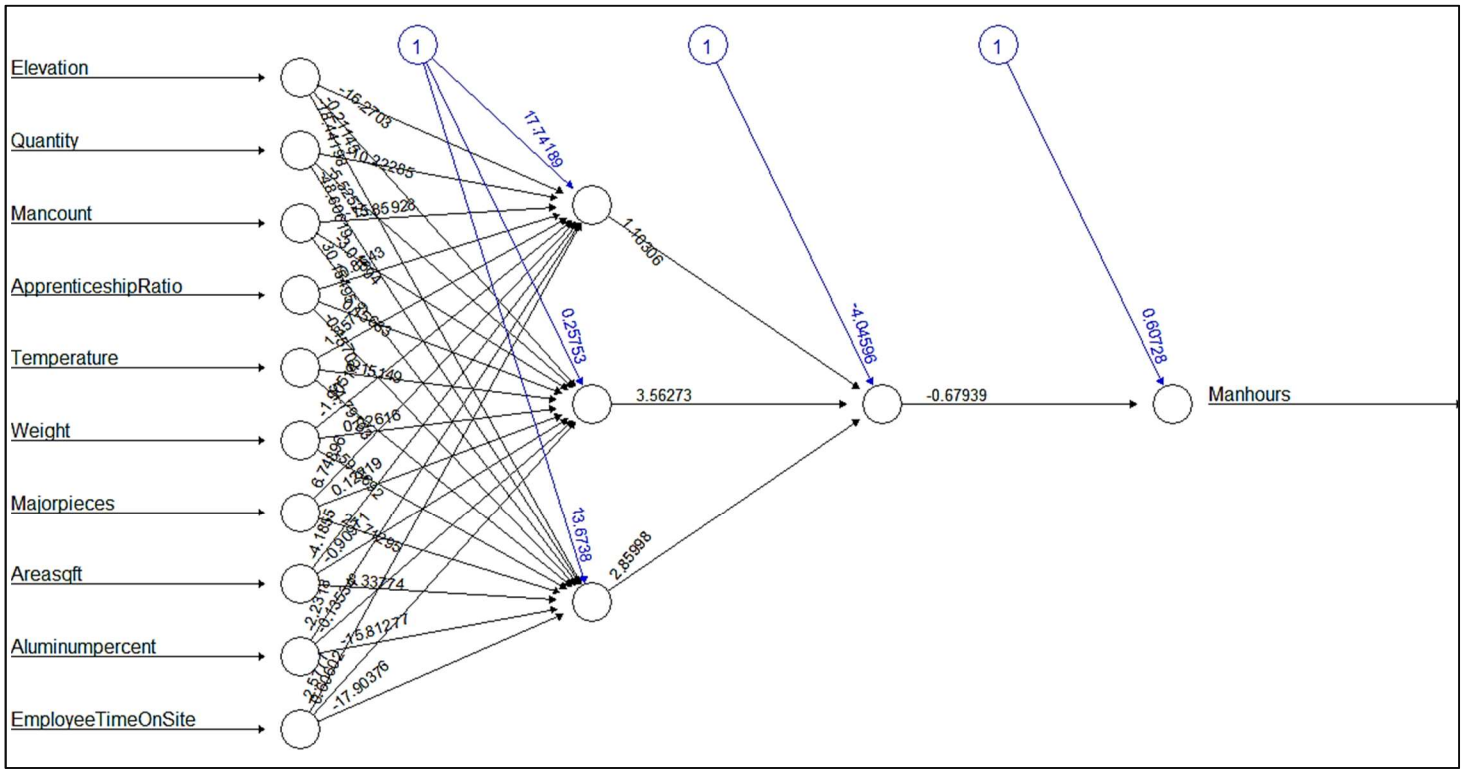
Appendix 1: Sample code for predictive model –Multiple Linear Regression

```
Call:
lm(formula = Manhours ~ Quantity + Weight + Elevation + Temperature +
    Aluminumpercent + Areasqft + Mancount + Majorpieces + I(Weight^2) +
    I(Majorpieces^2) + I(Quantity^2) + I(Elevation^2) + I(Mancount^2) +
    I(Temperature^2) + I(Areasqft^2) + I(Aluminumpercent^2),
    data = trainingset)

Residuals:
       Min         1Q     Median         3Q        Max
-716.98149  -28.97308   -4.96137   20.14007 1187.13207

Coefficients:
                             Estimate         Std. Error  t value
(Intercept)          -19.70560470137021   4.79019295611875 -4.11374
Quantity               0.25996870329789   0.01679276474970 15.48099
Weight                -0.00080781484898   0.00194552868136 -0.41522
Elevation              1.19368671542467   0.41609145842691  2.86881
Temperature           -0.79104167945063   0.12468902853296 -6.34412
Aluminumpercent       57.65294162390348  31.29943290836548  1.84198
Areasqft               0.03114424434828   0.00920878082973  3.38202
Mancount               4.66472265619990   0.43441831341098 10.73786
Majorpieces            0.00844393223113   0.02754774937513  0.30652
I(Weight^2)           -0.00000004168538   0.00000002913078 -1.43097
I(Majorpieces^2)      -0.00002346819065   0.00001007135286 -2.33019
I(Quantity^2)          0.00001172262530   0.00000359517178  3.26066
I(Elevation^2)        -0.00866589758363   0.00991894076096 -0.87367
I(Mancount^2)         -0.02275586812338   0.00582297046158 -3.90795
I(Temperature^2)       0.02946133774668   0.00870839876476  3.38309
I(Areasqft^2)          0.00000286662679   0.00000123293006  2.32505
I (Aluminumpercent^2) -52.59843549983992  46.41147406223499 -1.13331
                                Pr(>|t|)
(Intercept)                 0.00004006343239 ***
Quantity                < 0.000000000000000222 ***
Weight                             0.67801583
Elevation                          0.00415137 **
Temperature                 0.00000000026037 ***
Aluminumpercent                    0.06558474 .
Areasqft                           0.00072954 ***
Mancount                < 0.000000000000000222 ***
Majorpieces                        0.75923187
I(Weight^2)                        0.15255034
I(Majorpieces^2)                   0.01986738 *
I(Quantity^2)                      0.00112499 **
I(Elevation^2)                     0.38237273
I(Mancount^2)               0.00009530971457 ***
I(Temperature^2)                   0.00072669 ***
I(Areasqft^2)                      0.02014100 *
I(Aluminumpercent^2)               0.25718344
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 48.14588 on 2773 degrees of freedom
Multiple R-squared:  0.8961036,       Adjusted R-squared:  0.8952733
```

Appendix 2: Sample code for data analysis – Random Forest Regression

```
Call:
 randomForest(formula = Manhours ~ Quantity + Weight + Elevation +
Temperature + Aluminumpercent + Areasqft + Mancount + Majorpieces +
 I(Weight^2) + I(Majorpieces^2) + I(Quantity^2) + I(Elevation^2) +
I(Mancount^2) + I(Temperature^2) + I(Areasqft^2) + I(Aluminumpercent^2),
data = trainingset, importance = TRUE, ntree = 500)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 5

          Mean of squared residuals: 7214.074
                    % Var explained: 82.9
```

Appendix 3: Neural network generated by R studio



Appendix 4: Actual and Predicted Values of regression algorithms.

Actual Values	Linear regression	Random Forest	Decision tree
156.18	163.145	241.91	105.43
51.16	22.325	39.405	36.93
293.54	145.785	217.13	105.43
187.86	89.925	81.28	105.43
17.61	5.77	16.78	36.93
2152.47	1333.705	982.045	1150.31
392.91	199.425	174.265	154.075
47.36	18.705	23.965	36.93
79.62	55.78	60.12	36.93
7.81	2.76	11.585	36.93
46.08	20.89	20.295	36.93
259.59	143.685	121.96	145.475
85.27	40.61	37.33	36.93
129.82	71.025	82.215	36.93
78.5	36.23	44.015	36.93
285.69	115.075	127.41	105.43
47.6	25.105	35.765	36.93
112.78	61.405	75.645	36.93
19.33	20.83	26.54	36.93
88.51	40.115	119.68	105.43
45.04	18.18	15.12	36.93
70.38	38.725	64.795	36.93
167.2	81.06	106.97	105.43
129.41	63.89	123.435	105.43
45.38	21.82	18.24	36.93
25.21	7.975	17.455	36.93
143.13	84.01	188.995	105.43
96.44	44.315	65.19	36.93
64.29	35.075	39.21	36.93
64.08	32.31	52.33	36.93
13.39	15.09	35.465	36.93
22.89	9.765	29.18	36.93
28.94	18.665	17.505	36.93
69.94	33.88	49.69	36.93
190.52	96.565	73.14	36.93
148.34	70.46	73.995	105.43
81.99	39.39	62.425	36.93
12.86	7.46	26.735	36.93
144.7	48.13	39.175	36.93
122.4	59.605	59.255	36.93
192.93	91.93	158.63	254.075

23.28	26.075	30.035	36.93
47.64	20.475	17.585	36.93
13.61	7.225	24.65	36.93
25.52	9.34	18.625	36.93
15.31	7.815	16.04	36.93
84.22	52.39	68.125	36.93
56.14	30.99	32.135	36.93
91.87	46.15	46.22	36.93
41.61	9.14	12.47	36.93
14.86	18.55	21.895	36.93
22.29	20.015	15.635	36.93
84.48	70.78	79.205	105.43
25.6	19.035	17.495	36.93
1664.86	827.19	891.145	541.4
118.17	65.755	92.325	105.43
343.89	177.67	152.49	254.075

Appendix 5: Actual and Predicted Values of artificial neural network algorithms.

Actual Values	Neural networks
14.56	11.975
20.09	21.33
3.83	2.105
25.6	19.83
25.6	19.15
8.585	0.975
40.74	50.175
39.39	32.72
474.34	473.11
14.825	12.23
15.12	18.67
24.48	32.385
46.83	49.475
30.37	23.785
13.665	14.305
2.905	8.285
8.72	9.86
17.165	15.22
26.815	33.61
47.375	39.245
71.51	77.42
25.585	20.705

26.815	33.995
28.54	31.045
28.54	33.26
24.63	15.215
38.86	36.575
22.935	31.665
21.995	14.17
46.925	50.825
37.165	33.625
32.255	40.14
15.05	21.135
47.485	40.225
16.515	19.285
25.81	30.44
75.705	82.345
59.53	66.79
66.755	62.55
52.52	55.42
7.685	17.465
42.27	42.45
22.025	23.975
59.19	68.42
24.875	21.435
32.335	28.985
81.97	73.925
12.435	17.23
94.6	93.175

Appendix 6: Cumulative Results of multiple iterations done based on grouping the data sets.

Data Set	Algorithms	Test Results		
		RMSE	MAE	R squared %
Complete Data Set	Multiple Linear Regression	105.75	53.58	62.3
	Random Forest	111.91	47.44	64.73
	Decision Tree Regression	129.17	63.68	58.05
	Neural Network	119.21	49.91	62.59
Data Set grouped by dividing data into different subsets based on Work Classification	Multiple Linear Regression	69.3	43.8	83.6
	Random Forest	107.93	50.01	83.85
	Decision Tree Regression	145.71	78.85	68.06
	Neural Network	114.53	53.60	73.6
Data Set grouped by dividing data into different subsets based on Work Classification and disciplines.	Multiple Linear Regression	45.08	25.48	90.1
	Random Forest	73.14	43.98	87.26
	Decision Tree Regression	103.8	62.45	75.43
	Neural Network	15.25	25.68	91.2