Application of Data Mining Techniques for Conceptual Cost Estimation of Selected Building Projects in Addis Ababa.

A thesis conducted in partial fulfilment of the requirements for

Masters of Science in Civil Engineering

(Construction Technology and Management Major)

By

Biruk Lemlem Adam

ADDIS ABABA UNIVERSITY

ADDIS ABABA INSTITUTE OF TECHNOLOGY

School of Civil and Environmental Engineering

Advisor: Dr. Abraham Assefa Tsehayae

Addis Ababa University

Addis Ababa Institute of Technology

School of Civil and Environmental Engineering

Application of Data Mining Techniques for Conceptual Cost Estimation of Selected Building Projects in Addis Ababa.

Biruk Lemlem Adam

Approved by Board of Examiners:

Dr. Abraham Assefa Tsehayae	for Adm	May 26th, 2022
Advisor	**Signature**	**Date**
Dr. Gebrehana Tadesse	Gebre.	19 May 2022
External Examiner	**Signature**	**Date**
Selam Yazew		June 2, 2022
Internal Examiner	**Signature**	**Date**
	Mebruk Mohammed (Dr.-Ing.) Dean, School of Civil & Environmental Engineering	
Chairman	**Signature**	**Date**

Author's Declaration

I hereby declare that the work which is being presented in this thesis entitled "***Application of Data Mining for Conceptual Cost Estimation of Selected Building Projects in Addis Ababa***" is the original work of my own, has not been presented for a degree in any other university and all the resources of materials used for the thesis have been duly acknowledged.

Biruk Lemlem Adam **Date**

(Candidate)

This is to certify that the above declaration made by the candidate is correct to the best of my knowledge.

Dr. Abraham Assefa Tsehayae **Date**

(Thesis Advisor)

ACKNOWLEDGEMENT

I would like to begin by praising GOD, the almighty, for he is always watching over me. His love and grace is the reason I am able to complete this research and I believe his guidance, along with my hard work will only lead me into a life full of success, health and blessings.

I would like to thank my advisor, Dr. Abraham Assefa Tsehayae, for his invaluable guidance not just during this research but throughout the entire graduate program. His dedication towards modernizing Ethiopia's construction industry which he has exhibited through his exciting lectures as well as his patience and work ethics to read through my drafts and weed out and correct even the slightest mistakes I have made is not wasted on me. Thank You!

I would also like to thank my family and friends for their support throughout my graduate studies, especially my recently departed grandmother, W/ro Gebeyanesh Teklemichael, who was supporting and encouraging me with her wise words and prayers up until the completion of this research, I love you and may God rest your soul in heavens!

Finally, I would like to acknowledge the people from the Addis Ababa construction bureau, Ethiopian construction design and supervision works corporation, Ethiopian defense construction enterprise as well as the private consultants and contractors who were willing to provide me their data. This research wouldn't have been done without your support and willingness.

Thank you!

ABSTRACT

For project managers and decision makers, developing an accurate cost estimate in the conceptual stage of a project is a crucial but challenging task. Different techniques and methods have been devised and researched to accurately estimate the cost of building projects at the preliminary stages. These methods can broadly be divided into two based on the approach they follow. The cost –based or parametric cost modeling approach uses historical cost data and different Data Mining techniques to develop a cost prediction model. The second method uses a bottom-up or quantity strategy, in which data on the quantity of works is utilized to construct quantity prediction models for each work item. These predicted quantities can then be multiplied by their current unit rates to determine the respective costs. In this study a parametric cost model is first developed to assess its accuracy in predicting the final cost of building projects based on historical data collected from selected building projects in Addis Ababa. This was then followed by doing a comparison between the cost based and quantity based approaches by developing models for structural cost prediction as well as quantity models for the different work items that make up the structural work (concrete, reinforcement, and formwork). Concurrently, the study explored the effectiveness of four data mining techniques, namely Linear Regression (LR), Decision Trees (DT), Neural Networks (ANN), and Gradient Boosted Trees (GBT) in estimating the final and structural cost of building projects. With a relative error of 37.05%, the ANN model was the most accurate in forecasting the final cost of a construction project, while the GBT model performed better in predicting structural costs with a relative error of 22.67%. For quantity estimation models, the NN model showed superior performance for concrete and reinforcement quantity estimation with a relative errors of 16.44% and 19.32% respectively. The GBT model on the other hand performed better in formwork quantity estimation with a relative error of 19.58%. Accordingly, the total slab area was identified to be the most important variable for all prediction. The study indicated the quantity based approach provides more accurate cost prediction as opposed to the cost based approach for the case of structural cost estimation.

Keywords: Data mining, Conceptual Cost Estimation, Structural Cost Estimation, Quantity Estimation, Artificial Neural Network, Gradient Boosted Trees, Linear Regression, Decision Trees

LIST OF ABBREVIATIONS

- **AE** – Absolute Error
- **AFH** – Average Floor height
- **AHP** – Analytical Hierarchical Process
- **ANFIS** – Adaptive Neuro-fuzzy System
- **ANN** – Artificial Neural Network
- **BF** – Basement Floors
- **BP** – Back Propagation
- **BT** – Building Type
- **CBR** – Case based Reasoning
- **CMQ** – Construction Material Quantity
- **CRISP**-DM - Cross-industry standard platform for data mining
- **CV** - Cross validation
- **DS** – Data System
- **DT** – Decision Trees
- **ED** – External Decoration
- **EFNIM** - Evolutionary Fuzzy Neural Inference Model
- **FALCON** – Fuzzy Adaptive learning Control Network
- **FS** – Fire Alarm System
- **FT** – Foundation Type
- **GA** – Genetic Algorithm
- **GBT** – Gradient Boosted Trees
- **GNR** – Generator
- **IDF** – Internal Decoration (Floor)

- **IDW** – Internal Decoration (Wall)
- **KDD** – Knowledge Discovery in Database
- **LR** – Linear Regression
- **MAPE/RE** – Mean Absolute Percentage Error/Relative Error
- **ML** – Machine Learning
- **MLP** – Multilayer Perceptron
- **MSA** – Multi Step Ahead
- **NF** – Number of Floors
- **NL** – Number of Lifts
- **PCA** – Principal Component Analysis
- **RC** – Reinforced Concrete
- **RMSE** – Root Mean Squared Error
- **SLT** – Slab Thickness
- **SSE** – Sum of Squared Error
- **ST** – Slab Type
- **SVM** – Support Vector Machine
- **SVR** – Support Vector Regression
- **SW** – Shoring Work
- **TSA** – Total Slab Area
- **XGBoost** – Extreme Gradient Boosting

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

As Civil Engineers, we are usually asked by prospective clients or friends/colleagues how much it would cost to build a certain type of building with certain size and features. Traditionally we answer these questions by basing on a certain value per unit area of the building. This value is often dependent on past experiences and mostly, the estimate we provide is far off the accurate one.

Quite recently the Addis Ababa city government has announced its plan to build 500,000 houses in and around the city. While this plan is an ambitious one, in a recent lecture, a concern was raised regarding whether the necessary cost estimation was done to answer how much money such projects will require and if there is sufficient supply of construction materials in and around the city to even entertain such a big undertaking. This brought on the question: is there any accurate way of estimating the cost and required materials for a project during the early stages before even the design for the building has commenced? This was the motivation behind this research topic.

Before investing on a project, a corporation or organization's first step is to conduct a financial and technical assessment of the proposed project. While the technical analysis may take a variety of approaches depending on the project type, the financial analysis follows a similar pattern: Identify a project or projects, determine cost of the project, estimate the benefits, perform cost-benefit analysis and chose the one with highest benefits.

Construction projects, just like any other businesses, require to be studied to identify whether they are feasible or not. While the benefits can be subjective and may have different meanings for private and government projects, one thing that relates all is the cost. In the early-stage, cost estimates allow businesses to compare between projects, change the scope of the projects according to their budget, secure funding and use these estimates as a basis for cost control during the construction period. The use of poor strategies to estimate the cost of projects at the preliminary stage can easily turn an expected profit into a loss (Elbetagi, 2015). This is where conceptual or preliminary cost estimating comes in.

Conceptual cost estimating is defined as the forecast of project costs that is performed before any significant amount of information is available from detailed design and with incomplete

work scope definition. The main purpose of conceptual estimate is to be a basis for important project decisions like whether to proceed with the project or not and when it is decided to proceed with the project, in the appropriation of funds decisions (Elbetagi, 2015). Often in public projects, a conceptual cost estimate is also used to set a preliminary construction budget, and to control construction costs during the design.

Despite the importance of early cost estimates, arriving at one is not a simple and straightforward task. This is mainly credited to the lack of information that is available during this early stage of the project. As a result, these estimates significantly vary from the actual construction costs. Lack of accurate preliminary estimates has been linked with cost overruns in different literatures mainly (Nyoni, 2019; Zinabu, 2015; Bekele, 2017 and Samuel, 2014). To improve the accuracy of these estimates, different researchers have attempted to develop models that predict the cost of construction projects at the preliminary stage. These attempts can broadly be classified into two depending on the approach taken.

The first and famous method is to collect historical cost data of buildings and develop cost estimation models, be it parametric like linear regression or non-parametric like Artificial Neural Network, Case based reasoning, Fuzzy Logic, Support Vector Regression, etc. This approach is referred to as the *Cost-based approach or Cost modeling*. In many occasions, early cost estimations at the conceptual phase using this approach and techniques are reasonably accurate for early estimates. Despite their satisfactory accuracies, these models were found to depend significantly on the prices of previous projects to estimate the cost for the future ones. The problem with these estimation as noted by Joseph (2013) is they need a somewhat stable market condition where the costs for materials, and labor can be reasonably forecasted and the inflation, well represented by cost indices. Zhang (2017) also points out that in areas where there is high price fluctuation, the estimates from these models will be far less accurate. Unfortunately, Price fluctuation in the Ethiopia occurs in an unpredictable manner with increase in material prices by more than 34% - according to a study done by Asteway (2008).

In order to address this issue, the second method or approach being taken by researchers is to use historical quantity of works data from previous project and to develop quantity prediction models using parametric and/or non-parametric techniques. Once the quantity for the different work items is known, a current unit price of each work item can be used to calculate the

respective costs (Borja & Bryan, 2016). This approach is known as the *Quantity – based approach*. An illustration for these approaches are provided in Figure 1.

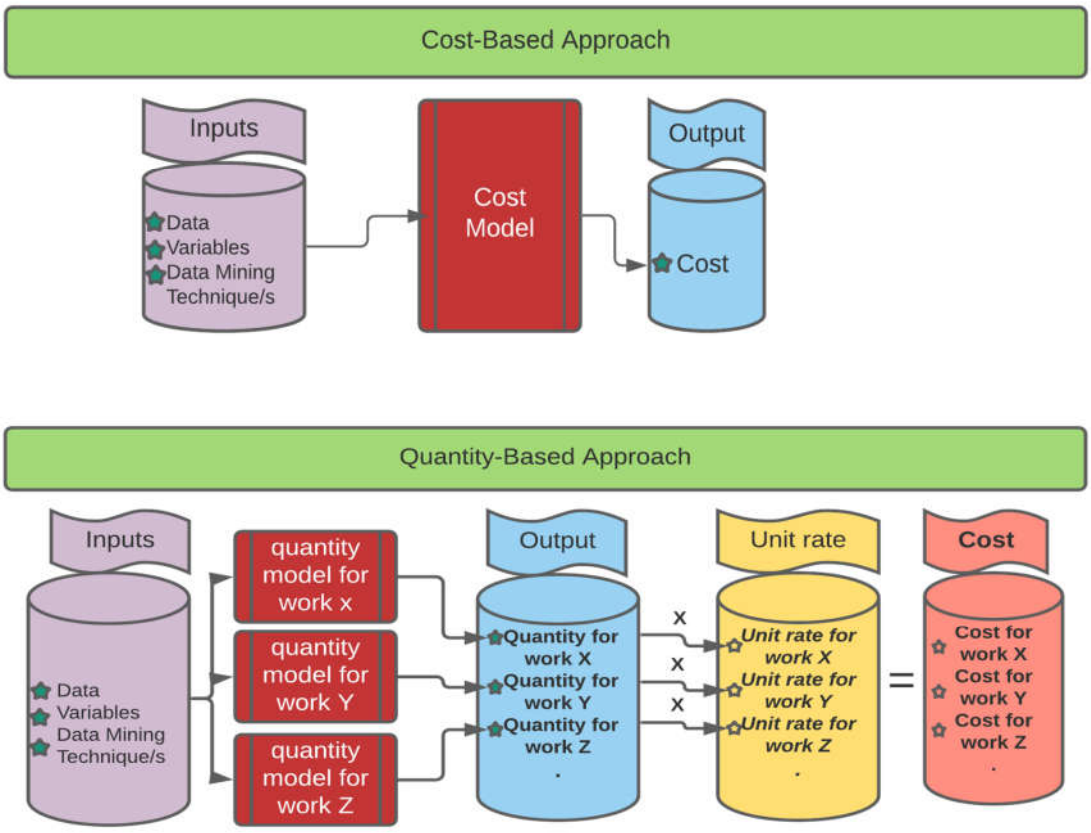

Figure 1: Illustration of Cost-based and Quantity-based Approach

Data Mining is defined as "a multi-disciplinary skill that draws upon machine learning, Statistics and database technology, information retrieval and Artificial Intelligence to look for useful patterns in data sets" (Sumathi & Sivanandam, 2006).

The researches discussed above as well as in the literature review tend to use statistics and machine learning techniques to look for hidden patterns in cost and quantity data of previously completed projects to develop prediction models, thus, by definition, falling into the data mining category. This study attempts to leverage data from selected building projects in Addis Ababa and some of the well-known data mining techniques to develop final cost estimation models as well as to compare the accuracy among these two approaches for structural cost estimation of building projects in Addis Ababa.

1.1 PROBLEM STATEMENT

The Problems Outlined for this research is two-fold;

Primarily, given most clients, developers or government bodies are working within tight, pre-defined budgets, they expect to know the cost of a project before investing more time and money into it. This requires engineers to come up with an accurate preliminary/conceptual cost estimates so that better informed decisions can be made, the projects can be well planned and any cost overrun as a result of poor estimation mitigated. This need for accurate preliminary cost estimations have been outlined in different literatures: (Samuel, 2014; Zinabu, 2015; Bekele, 2017 & Tadesse, 2018).

While various researches have been done in this area abroad, most of the researches concerning cost modeling in Ethiopia have focused on road projects. Whereas studies by Alemu (2020) and Abebe (2017) has focused on modeling bid markup estimation for road projects using ANN and LR, as well as on determining optimum cost contingency using Monte Carlo Simulation, respectively, It is Alemayehu (2014) who first developed Linear regression models to estimate the preliminary final costs of road projects, which was later followed by Tadesse and Dinku (2017), who developed and compared preliminary cost estimation models using Linear Regression and Artificial Neural Networks.

Beside the aforementioned researches, there has been no published research attempts made in adopting or comparing different data mining techniques in estimating the conceptual/ preliminary cost of building projects in the country – indicating a research gap in the area.

The second point is, besides the issue discussed in the introduction part, these cost – based models have another fundamental shortcoming: this is, the estimates gained from these models only provide a single monetary value and does not really give any information on the scale of work to be expected or the quantity of materials needed for the project. The availability of such information would allow parties involved in planning such projects to check for availability of the materials as well as equipment in the market and if shortage for these materials exist, to plan ahead for it - a notion agreed by Oluwafunmibi and Lam (2020) and García de Soto (2014).

A potential alternative for this is the quantity-based approach. Instead of developing a cost model that predicts the final cost a building, one can break down the project into different work

items (as done in quantity surveying) and apply data mining techniques to develop a conceptual quantity prediction model for each work items. Once the quantities are estimated, one can use a current (latest) unit rate (price) for each work items and determine the costs of the individual work items that makeup the final cost. This would provide clients with a rather clearer relationship between the different works and their corresponding costs. It is also useful in cost planning by determining which works require more capital and when these works will be carried out. In theory, this would also remove any cost fluctuation in the data from affecting the prediction models as instead of historical cost data, quantity data will be used to develop the models. This is especially important in countries like Ethiopia where there are high-cost fluctuations which can result in poor performance of cost (Zhang, 2017).

Quantity modeling has been researched abroad by different researchers. Yeh (1998) developed quantity prediction models for structural works of steel and composite building projects and concluded it provided more accurate estimates when compared to the cost-based approach. Singh (1990) developed a computer-based cost model to estimate the cost of reinforced concrete beam and slab construction in high-rise commercial buildings based on the quantity of works. Son et al. (2013) used the wall-to-floor ratio as the only predictor variable to predict quantities for concrete, reinforcement, and formwork in basement, ground, and upper floors and more recently, Oluwafunmibi and Lam (2020) used Support Vector Regression to develop models for predicting the quantities of reinforced concrete structural elements that can predict with an accuracy interval of 95%. Unfortunately, none of these techniques and approaches have been tested and adopted for the case of Ethiopia's building projects – indicating a gap in research in this area.

1.2. OBJECTIVE

1.2.1. MAIN OBJECTIVES

The main goal of this study is to address the aforementioned shortcomings by using the power of some well-known predictive algorithms to build a cost model to predict the final construction cost of building projects based on data from previously completed projects in Addis Ababa.

In addition, this study aims to compare the quantity-based and cost-based conceptual cost estimation approaches by developing various models using different data mining techniques for the structural work cost of these buildings.

1.2.2. SPECIFIC OBJECTIVES

The following specific goals have been planned in order to attain the main objectives:

- Identify and rank the main factors that influence the final and structural cost estimates, as well as the quantity of different structural works namely, concrete work, reinforcement, and formwork, at the conceptual stage.
- Evaluate different techniques of data mining, namely regression, decision trees, artificial neural networks and gradient-boosted trees, and evaluate their accuracy in the multiple predictive models to be built namely;
    - Final and structural cost estimation models and
    - Quantity estimation models for structural work of the buildings, particularly, the concrete work, formwork and rebar.
- To compare the accuracy between cost-based and quantity-based approaches in estimating the structural cost of building projects and identify which approach (cost-based or quantity-based) provide more accurate estimations for structural cost estimation

1.3. SIGNIFICANCE OF THE STUDY

This study is expected to have significance both for practitioners and researchers:

- For the practitioner, this research will provide them with the methods on how to develop a tool to estimate the cost of building projects at the preliminary stage. Consultants, clients and even contractors can use this tool to estimate, at the early stage, how much it will cost them to build buildings.

  Furthermore, this research can be a guideline on how to use their past data and a free-to-use data mining software on developing different models and use the quantity-based cost estimation approach to determine the quantity of different construction works and materials required for the structural and other works of building projects. Having this knowledge at the preliminary stage will aid developers in identifying if/whether there is enough supply of the material on the market and if there is shortage, to plan ahead for it.

- For the researcher, this study is expected to fill the research gap seen in the country regarding preliminary cost estimation of building projects. Furthermore, it is expected to be a great addition to other researches being done in this area. It will also help researchers in identifying if/weather the unstable market situation seen in our country has an effect in the preliminary cost estimation of building projects based on past cost data and if it is not, what other approach can be used to get more accurate predictions.

2. LITERATURE REVIEW

2.1. SECTION I – COST ESTIMATION

2.1.1. COST ESTIMATION: DEFINITION

Searching the definition for the term "cost estimation" presents more or less similar results that basically say, "Cost Estimation can generally be defined as a process of predicting the expenses of a project". This definition gives us the gist of what cost estimation is and though it hints the complexity of it by defining it as a process, it lacks to describe what exactly the process entails.

In an effort to understand the term even more, perhaps it would be better to first define the two words that make up the term – 'Cost' and 'Estimation'. From project management perspective**,** a cost can be defined as the amount of money necessary to accomplish a certain project; a means to an end, one can say, where the end is achieving the project goals. The mathematical definition of Estimation is, "The approximation of the value of a population parameter on the basis of sample statistic." (Niu, 2020). In cost estimation case, the parameter is the cost whereas in quantity estimation, the quantity becomes the parameter.

The Project Management Body of Knowledge Guide (PMBOK® Guide) provides a more distilled and specific definition of estimating as, "A quantitative assessment of the likely amount or outcome. Usually applied to project costs, resources, effort and duration and is usually preceded by a modifier (i.e., preliminary, conceptual, feasibility, order-of-magnitude, definitive). It should always include some indication of accuracy (e.g. ±x percent)" (PMI, 2017).

Chen and Richard (2002) on the other hand, defined estimating as, "a complex processes that involves collection of available and useful information, taking this information and visualizing each process of the projects' and ultimately translating this visualization to the final cost."

From these two definitions, the following points can be outlined about cost estimation:

- *Cost estimation is a process and a cost estimate, the outcome*
- *A cost estimate can be subjective to the estimator's judgement and amount of information available.*

- *There are different types of cost estimates and cost estimate is never 100% accurate.*

With that in mind, a more inclusive definition of cost estimation is one presented by Wrike (2020) that states,

> *"A cost estimation is the process of forecasting the financial and other resources needed to complete a project within a defined scope. Cost estimation accounts for each element required for the project—from materials to labor—and calculates a total amount that determines a project's budget."*

The Project management Institute also have somewhat similar definition;

> *"Cost estimation is the predictive process used to quantify cost and price of the resources required by the scope of an investment option, activity or project and it is a process used to predict uncertain future costs with a goal of minimizing the uncertainty of the estimate given the level and quality of scope definition. Cost estimation implies a process of assessing and predicting the likely cost of resources (labor, material, time) needed for completion of a certain project."* (PMI, 2017).

Importance of Cost Estimating

The main purpose of cost estimates is well, to determine the cost of a project. But what do we benefit from determining these costs? Some of the benefits associated with determining these costs include:

- To make early decisions regarding the feasibility of the project
- To reduce or increase the scope of the project based on budget
- To find and secure funding for the project
- To plan more accurately
- To manage resources (labor, equipment and money) properly
- To manage risks

2.1.2. METHODS AND TYPES OF COST ESTIMATES

There are different types of cost estimates based on the requirements of the project and the stage the project is at the time. More often than not, the methods used for these estimates dictate the type of the estimate and vice versa. Thus, the major methods and types of cost estimates particularly used in construction projects are explained below.

The main types of estimates based on the stage of the project can be summarized as follows:

- Conceptual cost estimates,
- Semi-detailed cost estimates and
- Detailed Cost estimates.

A conceptual estimate is also known as a top-down, order of magnitude, feasibility, analogous, or preliminary estimate. It is the first serious effort made to predict the cost of the project. A conceptual estimate is usually performed as part of the project feasibility analysis at the beginning of the project. There is shortage of information regarding the projects' design at this stage and estimates are usually based on verbal or written description of the projects. As a result, it has an error of ± 25% (Elbetagi, 2015).

Semi-detailed cost estimates are developed while basic design decisions are being made to verify that the project can be constructed at its intended scope within the owner's budget. Some aspects of the project may be completely designed. These estimates are expected to be within ± 15% of the actual cost (Elbetagi, 2015).

A detailed estimate is also known as a bottom-up, fair-cost, or bid estimate. Detailed estimates are prepared once the design has been completed and all construction documents prepared. The estimator divides the project into individual elements of work and estimates the quantities of work for each element. Since the design is completed at this stage, it is expected to have an error less than ± 5% (Elbetagi, 2015).

Cost estimates on the other hand can also be grouped based on the required accuracy and speed, which also dictate the methods/techniques. These are:

1. Quantity survey estimates: This estimate involves detailed calculation of the quantities and costs required to complete the project. The estimator will breakdown the project into

different activities and calculate the required manpower, materials and other indirect costs required to complete each activity. The final cost will be the final sum of these activity estimates.

Such estimates require the estimators do a quantity estimation, work breakdown schedule as well as estimate of indirect costs prior to estimating the costs. These estimates take too much time to complete depending on the complexity of the project at hand. Thankfully modern technologies, like Building information modeling (BIM), have made this easier by automatically determining the quantity of works or materials from the designs.

2. Unit price-based cost estimates: these estimates are prepared by using a cost per unit square area and multiplying it with the total area of the proposed project. The cost for a unit square area can be found based on previous, similar projects or using guides that suggest ranges of costs for different building types.
3. Parametric estimates: this technique requires the use of past building cost data to form a statistical relationship between building parameters, hence the name parametric estimation. The parameters used may include floor heights, gross or floor areas, number of floors and so forth.
4. Model based estimates: These estimates are types of parametric estimates that base their estimate on previously built computer models that estimate the cost of a project. These models require the user to answer few questions regarding the important characteristics of proposed buildings and based on that and depending on which statistical or machine learning algorithms the models are built on, they will provide the user with an estimate. The upside of such estimates is they can provide conceptual estimates or detailed ones based on how these models are built. Unfortunately, these models are not universal and need considerable effort and data to be built so they provide accurate results.
5. Expert Judgment estimates: Often used to determine conceptual cost estimates, this technique entails the use of specialist estimators to use their years of experience in estimating different projects to create an estimate for a project.

A table summary of the type of estimates along with which methods works for them is provided in table 1.

Table 1: types of estimates and methods adopted to do the estimations

	Quantity survey	Unit price	Parametric	Model based	Expert Judgment
Conceptual/Preliminary Estimates	√	√√√	√√√	√√√	√√√
Semi-detailed estimates	√	√√	√√	√√√	√√
Detailed Estimates	√√√	√	√	√√	√

Note: √√√ - Ideal Choices; √√ - can be chosen depending on availability of information; √ - Least choice

2.1.3. CLASSES OF ESTIMATES

Different institutions have adopted standards and estimate classes based on what the acceptable accuracy is for different estimate types. The most famous of these classes of estimate is the American Association of Civil Engineers classification system and is discussed below.

- ***Estimate Class 1***

Estimate Class 1, also called Full detail estimates, bottom-up estimates or definitive estimates, has a project scope definition above 50%. This type of estimate is also known as definitive estimate. They are generally prepared for discrete parts or sections of the total project rather than for the entire project. Class 1 estimates should have an error range of -3% to -10% on the low side and +3% to +15% on the high side. These estimates can be used for evaluating bids, claim negotiations or dispute resolutions (American Association of Civil Engineers (AACE), 2005).

- ***Estimate Class 2***

Estimate Class 2 has between 30% and 70% of complete project definition. They are also known as definitive estimate, and are generally prepared to form a detailed control baseline against which all project work is monitored in terms of cost and progress control. Class 2 estimates always involve a high degree of deterministic estimating methods. Class 2 estimates are prepared in great detail, and often involve tens of thousands of unit cost line items. For those areas of the project still undefined, an assumed level of detail takeoff (forced detail) may be developed to use as line items in the estimate instead of relying on factoring methods. For contractors, this class of estimate is often used as the "bid" estimate to establish contract value.

- ***Estimate Class 3***

Also known as scope, semi-detailed or budget estimate, Class 3 estimate is generally prepared to form the basis for budget authorization, appropriation, and/or funding. These estimates have error ranges of -10% to -20% on the low side, and +10% to +30% on the high side, depending on the technological complexity of the project. They are typically prepared to support full project funding requests, and become the first of the project phase "control estimate" against which all actual costs and resources will be monitored for variations to the budget. In some cases, these estimates are the last estimate required and could well form the only basis for cost/schedule control (American Association of Civil Engineers (AACE), 2005).

- ***Estimate Class 4***

Class 4 estimates are also known as preliminary or conceptual estimates and are generally prepared based on limited information and, as a result, have fairly wide accuracy ranges. Their Typical purpose includes detailed strategic planning, business development, project screening at more developed stages, alternative scheme analysis, confirmation of economic and/or technical feasibility, and preliminary budget approval or approval to proceed to next stage. Typical accuracy ranges for Class 4 estimates are -15% to -30% on the low side, and +20% to +50% on the high side.

- ***Estimate Class 5***

Class 5 estimates are generally prepared based on very limited information, and subsequently have wide accuracy ranges. As such, some companies and organizations have elected to determine that due to the inherent inaccuracies, such estimates cannot be classified in a conventional and systemic manner. Class 5 estimates, due to the requirements of end use, may be prepared within a very limited amount of time and with little effort expended— sometimes requiring less than an hour to prepare. Typical accuracy ranges for Class 5 estimates are - 20% to -50% on the low side, and +30% to +100% on the high side. They are also called Ballpark or prospect estimates (American Association of Civil Engineers (AACE), 2005).

2.1.4. CONCEPTUAL COST ESTIMATION

A "conceptual estimate" or also called Preliminary estimate is a class 4 estimate prepared by using engineering concepts and avoiding the counting of individual pieces. As the name implies, conceptual estimates are generally made in the early phases of a project, before construction drawings are completed, often before they hardly begin. The first function of a conceptual estimate is to tell the owner about the anticipated cost, thus presenting useful information for the owner in contemplating the project feasibility and further development.

According to Elbetagi (2015), the concept of visualizing the project process that is discussed by Chen and Liew (2002) in defining 'cost estimation' is something of an art. Whereas using cost data of previous works to estimate cost has a scientific backing to it. He concluded that conceptual cost estimation is a mix of an art and science.

Preliminary estimates assist the overall cost-control program by serving as the first check against the budget. It will indicate the cost overruns early enough for the project team to review the design for possible alternates. Since preliminary estimate is made prior to the completion of detailed design, the margin of error will be relatively large. Then, the larger contingency should be applied. The contingency varies with the amount of design information available and the extent of cost information obtainable from similar projects.

The Need for Fast and Accurate Conceptual Cost Estimates

The estimation of a project cost plays a key role in the success of a construction project. The process of cost estimating is crucial as it enables construction companies to determine what their direct costs will be and to provide a "bottom line" cost, below which it would not be economical for the work to be carried out. The ultimate goal of these estimates is to provide decision makers a reasonably accurate information at the time they need it. Overestimating or underestimating the cost of a project can lead the parties involved into making ill-informed decisions that will incur loss in profit and opportunities. In this section, the importance of fast and accurate conceptual estimates will be discussed citing different researches done in the area.

Accurate estimate and Cost overrun

What many projects all around the world have in common is cost overrun. It would not be an overstatement to identify cost overrun as a plague in the construction industry. But cost overrun

is not a disease; just a symptom developed as a result of different underlying conditions and one of the many diseases attributed to cost overrun is poor estimating during the planning stages (Dominic D, 2014). Nichols (2007) Assessed 13 of the then largest road projects in the United Kingdom and identified inaccurate cost estimation, inflation and unsatisfactory project scope definition as the biggest causes of cost overrun. Perhaps a more extensive and representative research was done by Flyvbjerg (2008) who studied a total of 258 projects from 20 different countries and concluded incorrect estimations at the time of deciding to build among the major reasons of cost overrun.

Moving to Africa, a study done in Nigeria in 2014 after collecting and studying from 190 buildings data, identified that low rise buildings were being underestimated by up to 18.87% while high-rise buildings were being over estimated by 23.43% during their preliminary cost estimation. The researcher goes further into concluding that the conceptual cost estimates prepared in Nigeria are far less accurate than the range tolerated in the construction industry and that this could explain the high incidence of cost overruns presently being experienced in the Nigerian construction industry (Samuel, 2014). Similar result was arrived at by Nyoni (2019), who conducted a study to determine the factors influencing cost overrun in Zimbabwe's construction industry and identified that poor estimation of original cost as the number one cause.

Different researches have also been done in Ethiopia with regard to this topic. Zinabu (2015) in a master's thesis he did regarding the Ethiopian construction industry has also noted that inaccurate cost estimation method and poor planning are among the factors contributing to cost overrun. A study on factors affecting cost overruns in housing construction in Addis Ketema Sub city housing development project also concluded cost underestimation and lack of financial planning were among the main factors affecting cost over runs (Bekele, 2017). Similar conclusion were also reached by Yohannes Tadesse, who concluded the Ethiopian construction works corporation should work on making better cost estimates and plan the projects accordingly (Tadesse, 2018).

Other researchers like Ayele (2019) and Wendmu (2018) that conducted studies regarding the causes of cost overrun in different construction projects and companies, however, did not outline any relation to support the notion that poor estimates are causes of cost overruns.

It may be worthwhile to note that there seems to be a lack of a standard reference point to which cost overrun is calculated on. Some researchers like Tadesse (2018), Zinabu (2015), Nyoni (2019) and Holm (2005) defined cost overrun as the difference in cost between the time the project was approved and the final cost of the project; whereas Bekele (2017), Ayele (2019) and Wendmu (2018) took on the definition, "Cost overrun is defined as the change in contract amount".

Perhaps, it may be fair to assume that the different conclusions they arrived at was as a result of the very definition of 'cost overrun' they base their research on. But as this is hard to prove without conducting an independent research, let's leave this in the bursting bubble and move on. A resolve, middle ground one may say, to this issue is presented by Ahiaga-Dagbui (2014) who suggested Even though theoretically, these conceptual estimates seem to have little purpose to the project once the project scope and design parameters change, it is still important to note that it is based on these estimates that funding is secured and thus, are still viable as a reference point provided that there is little or no scope change to the project down the line.

Gunaydn and Dogan (2004) make a similar conclusion and indicate that early-stage decisions have a major impact on the cost of a construction. As the client, knowing what their budget is and the cost of the project, will be able to make informed decisions on the different aspects of the building like, height of the building, floor area, type of finishing materials to be used, etc.

Therefore, there is a need to accurately determine the approximate cost of a project during its conceptual phase so that the feasibility of the project can be studied more accurately and clients can make informed decisions on the potential cost of their proposed project.

- ***"The sooner the better"***

In an ideal world, clients would want their contractors or consultants to provide them with most accurate estimates as fast as possible. Currently, the only way to provide the most accurate estimate is to do the detailed estimating. Detailed estimating requires the designs to be completed but when the client is only at the stage of entertaining ideas or evaluating business opportunities, doing the detailed estimate can be time consuming and expensive.

For instance, if a client wants to invest in a project and if the project has n design parameters and the client wants to evaluate their options by varying one parameter at a time, this would

generate (n 1)n different project alternatives when $n > 2$ and n alternatives when $n <= 2$. This means for 2 design parameters, 2 different detailed estimates are needed and for 3 design parameters, this comes to 8 alternatives and increases exponentially. Doing the detailed estimate for all these alternatives will be costly and take too much time - which leads to delays in decision.

- ***"A good Plan today is better than a perfect one tomorrow"***

In order to satisfy the client and deliver estimates on time, estimators' resort to faster techniques. Expectedly, this comes at a cost of accuracy. The accuracy of this conceptual estimate will depend greatly on the methods used and the availability of information at the time. So long as these estimates are somewhat correct, decision makers prefer to use these than spend more resources and wait weeks for a more accurate one. After all, "A good Plan today is better than a perfect plan next week" - George S. Patton (1885-1945). But here is another idea – why can't we make faster and yet, more accurate estimates? The entirety of this research can be concluded as finding ways to get more accurate and faster conceptual estimates by leveraging data on hand and some powerful data mining algorithms.

2.1.5. TRADITIONAL METHODS AND THEIR LIMITATIONS

While preparing the proposal for this thesis, an interview to different professionals residing in different consultancy offices in Addis Ababa was carried out in order to assess their current practices regarding estimating the cost of building projects at the conceptual stage – This was necessary as no published research was found on the current practices of conceptual cost estimation in Ethiopia.

Most of these companies base their estimates on the average unit cost method where the average cost per square meter of recently built projects is used to determine the conceptual cost of proposed projects. In order to accommodate for the differences in design between the proposed project and previously built one, adjustments are made based on expert judgment. This method is also called the superficial method. Literature suggests that this method is probably the most frequently used method of approximate estimating. Its major advantage is that most published cost data is expressed in this form. The method is quick and simple to use though, as in the case of the unit method, it is imperative to use data from similarly designed projects (Elbetagi, 2015).

Expert judgment is used extensively during the generation of cost estimates. Cost estimators have to make numerous assumptions and judgments about what they think a new product or project will cost. Even in those detailed estimates, expert judgment is still a valuable asset to have (Christopher, 2001). The problem with expert judgment is that it can be subjective and for conceptual estimates in particular, it is exposed to optimism bias. Optimism bias is "the tendency for people to be optimistic about future events, especially those seen as following from their own plans and actions. Although optimism is no doubt a stimulus to enterprise, it has obvious dangers: these include increased risk taking, failure to estimate probabilities accurately, and inadequate contingency planning. In drawing up plans, schedules, and budgets there is a demonstrated tendency for managers to underestimate costs and duration and to overestimate benefits." (Law, 2010).

Flyvbjerg (2008) outlined the concept of optimism bias in a paper published in 2008 where he suggested "decision making in infrastructure planning is susceptible to individual's optimism and blinds them from seeing the complexity of the process." This leads the estimators to become overly optimistic about a project that they will start underestimating costs and overestimating the benefits. Kahneman and Tversky (1979) also supports this notion. Their theory called the Prospect theory explains how people tend to make decisions without taking into account the actual outcome of their decision when they are presented with little information of the matter at hand. Optimism bias has become such a serious issue that in the United Kingdom, managers in government projects are expected to include an adjustment for optimism bias in their estimates (Law, 2010).

Several conventional conceptual cost estimation methods have also been used in the construction industry abroad. Methods such as average unit cost, cost indices and parametric estimation are among those used to rapidly calculate total project cost at the conceptual stage. These estimates also are heavily reliant on expert judgment and in many cases provide unsatisfactory results. To offset this problem, researchers and practitioners are always on the lookout for better techniques (Borja & Bryan, 2016).

Different quantitative techniques have been proposed and employed by researchers and practitioners for the estimation of construction costs at the early stage. While researchers and some clients and contractors use Machine learning techniques, such as Artificial neural networks (ANNs), Fuzzy logic, Support vector regression and case-based reasoning (CBR)

(with different variations), most practitioners still use models based on Statistical methods like regression analysis (García de Soto, 2014).

2.2. SECTION II - DATA MINING

2.2.1. INTRODUCTION TO DATA MINING

The traditional method of turning data into knowledge relies on manual analysis and interpretation. Even those using Expert judgments to estimate the cost, rely on past data or experience to extract some form of knowledge regarding construction costs. But this form of manual probing of data is slow and highly subjective as stated before. This along with the abundance of data available gave way to other, more sophisticated forms of identifying/ discovering knowledge from databases.

Data mining/KDD can be described as a process of identifying hidden, valid, and potentially useful patterns in data sets. It is a multi-disciplinary skill that uses machine learning, statistics, Artificial Intelligence and database technology to discover unsuspected/ previously unknown relationships amongst the data.

"The two high-level primary goals of data mining in practice tend to be prediction and description. Prediction involves using some variables or elds in the database to predict unknown or future values of other variables of interest, and description focuses on nding human-interpretable patterns that best describe the data. The goals of prediction and description can be achieved using a variety of particular data-mining methods" (Hamilton, 2018).

When the goal is to describe patterns or trends, the following tasks are generally adopted:

- Clustering: A process of partitioning the data sets in to similar classes.
- Summarizations: This technique is used to describe a rather large and complex data set in a simple, compact manner.
- Association Rules: are used to discover relationships between attributes and items such as the presence of one pattern implies the presence of another pattern. I.e. to what extent one item is related to another in terms of cause-and-effect. This is common in establishing a form of statistical relationships among different interdependent variables of a model
- Sequence analysis is used in a time-series data to identify sequential patterns.

Predictive Models/analysis on the other hand provides a forecast of what to be expected in the future. Predictive mining uses statistics and other machine learning algorithms to predict the future based on historical data. Methods such as Classification, Time series analysis and Regression fall under the predictive mining category.

Classification is the process of finding or discovering a model or function which helps in separating the data into multiple categorical classes. In classification data is categorized under different labels according to some parameters given input and then the labels are predicted for the data

Regression is the process of finding a model or function for distinguishing the data into continuous real values instead of using classes, as the classification. Regression can also identify the distribution movement depending on historical data. Because a regression predictive model predicts a quantitative value, it is expected to have more error than that of classification or descriptive models.

From the above definition, it can be said that estimation of construction costs based on historical data can be grouped in predictive data mining. But it is worth mentioning that some of the predictive models can be descriptive, to the degree that they are understandable, and vice versa, the distinction is useful for understanding the overall discovery goal. Thus, even those predictive data mining problems may adopt descriptive approaches depending on what exactly the goal is.

A summarized outline of data mining approaches and the most famous algorithms used are provided in figure 2

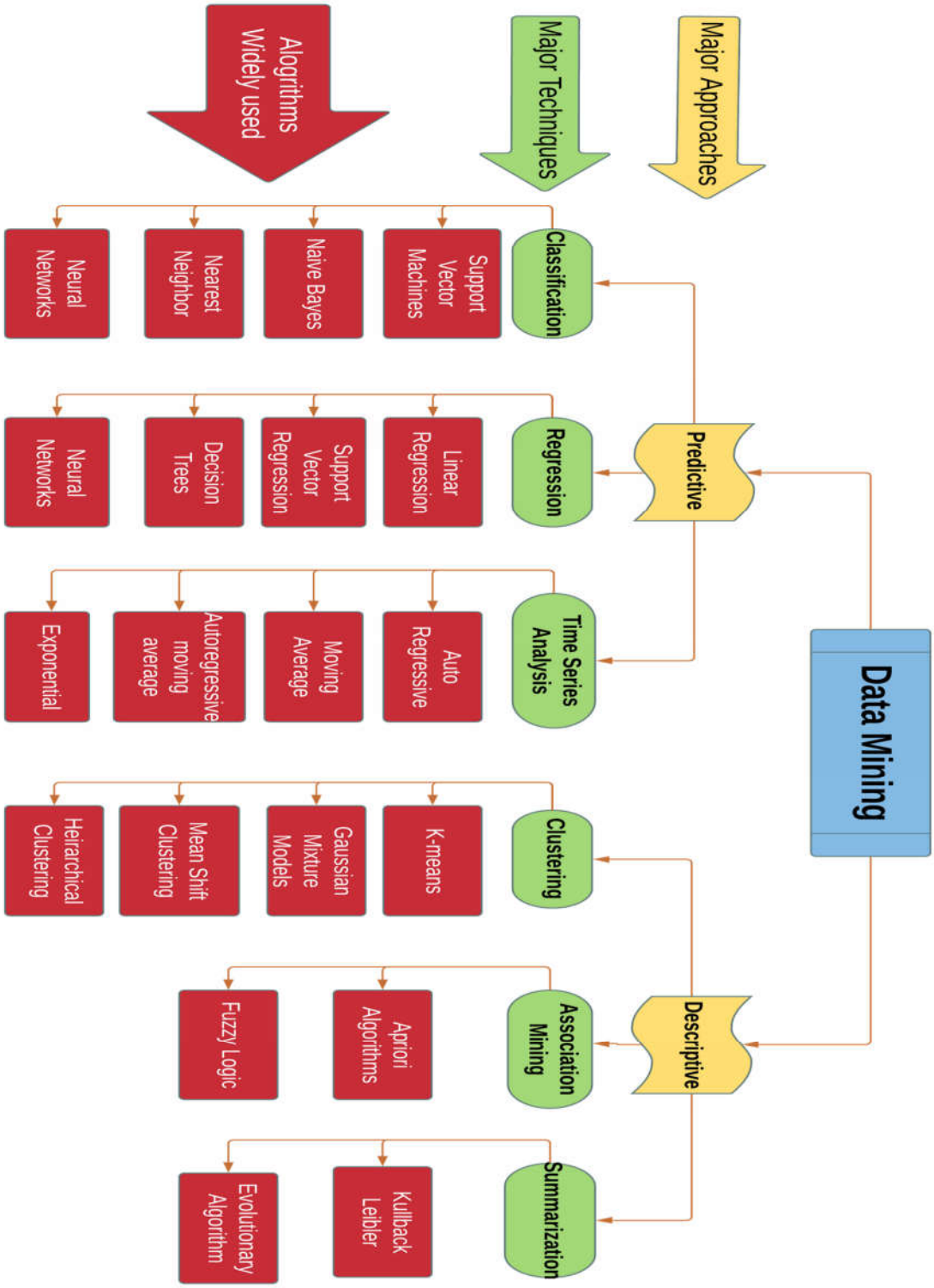

Figure 2: Data mining types and Algorithms

2.2.2. STATISTICAL METHODS

The regression analysis method is a statistical modeling approach that has been around for quite some time and has been used for all sorts of prediction modeling ranging from cost estimation to productivity of labor and equipment. According to García de Soto (2014), it is one of the most commonly used technique in statistical modeling.

Researchers like David, Margaret and Anthony (2016) have attempted to develop a linear regression model to predict construction cost of buildings. The research was based on 286 sets of building cost data that was collected in the UK. The researchers developed 6 models by varying the independent variables and the most accurate model had a mean absolute percentage error (MAPE) of 19.3% (David et al.,2006).

The Problem with regression Models for such problems as noted by Roxas and Ongpeng (2014) is the relationship between the input and output variables in regression models are straightforward and too simple compared to the complexity of the real-world relationship between those variables. Furthermore, regression models tend to perform badly when there is a presence of multi-collinearity - a statistical condition where there is a relationship between two or more independent variables that results in wrong predictions when using regression models.

In Ethiopia case, Alemayehu (2014) have attempted to develop and test regression models to estimate costs for Road Projects and managed to develop a model that estimates the final cost of road projects with a MAPE of 38.9% for the conceptual estimation. No published research was found for the case of building Projects.

2.2.3. MACHINE LEARNING METHODS

The study on the use of Machine Learning (ML) techniques for early cost estimation of construction projects is not a recent one. Many researches have been carried out since the 1980s to examine the applicability of ML methods like case-based reasoning (CBR), Support vector machine, Fuzzy logic and artificial neural networks (ANNs) in preliminary cost estimation of building projects. In the following sections, the application of some of these modern techniques in cost estimating will be discussed citing different researches.

2.2.3.1. CASE BASED REASONING.

Built based on the hypothesis that similar problems have similar solution, Case Based Reasoning (CBR) is a problem-solving technique that uses knowledge gained from previous experiences of the particular matter or cases and reuses it to solve a new problem (Roxas & Ongpeng, 2014).

Case based reasoning is among the first decision support system built next to rule-based systems. It has shown to be successful in different application domains, including but not limited to classification problems, planning and finding optimal solutions for different problems. In construction, case based reasoning has been studied with regards to its application in cost estimation as well as planning and scheduling (Koo et al., 2010).

The goal in CBR models is to find a case that matches the problem at hand. If the system can't find the same cases, it will find one that is most similar using case indexing methods like nearest neighbor methods which uses the square root of sum of the square of the arithmetic difference between two objects the case at hand and the case in the database in Euclidean space the distance between objects, inductive reasoning and combination of the two (Ji et al., 2011).

This system estimates the cost of new building projects by depending on the similarity between the proposed projects in its data base. Though this is an upgrade to expert judgment, which also is based on the expert's previous projects experiences, as it is faster, accurate and not susceptible to biases, it also has the core drawbacks of expert based estimates; just like experts with years of experience provide better estimates than those with few years of experience, the CBR performance is also dependent on the number of cases it is fed. Thus, it requires a significant amount of past building data to accurately estimate the cost for buildings (Brighterion, 2017).

Much recently, CBR was adopted to estimate the costs of a sports field construction. The researchers identified 7 variables with the highest correlation to the cost of the projects and developed their CBR model on them. Out of these 7 variables, two were related to sustainable development. The model was applied to estimate the cost of 3 sports fields and was able to do so with a MAPE of 14%. They concluded that, "In the short run, factors like impact of the object and the type of materials that are used from the perspective of their in uence on the

environment may be decisive as far as the costs determined in the life cycle of the building are concerned" (Lesniak & Zima, 2018).

The Inherent problem with case based reasoning is that the quality of result one can get from it is highly dependent on the number of cases it has fed the model. Zima (2015) who attempted to estimate the conceptual estimates of building projects in Poland concluded that though the result he found was satisfactory for the case of conceptual estimates, a larger data base is necessary if we are to use this as a decision support tool.

2.2.3.2. ARTIFICIAL NEURAL NETWORKS

2.2.3.2.1. BACKGROUND

Artificial Neural networks are arguably the most researched artificial intelligence techniques when it comes to conceptual cost estimations. Artificial Neural Network, as the name suggests, is an abstraction of the human brain with abilities to learn from experience and generalize based on acquired knowledge (Dominic D, 2014). Though artificial neural networks are not as complex as biological neurons, they are inspired by the same principle:

A neural network is composed of different layers each having their own nodes or neurons. These layers can be grouped as the Input layer, Hidden Layer and Output layer. The network takes the input value, which is a normalized value of the actual inputs, from the input layer and passes them through a hidden layer. The hidden layer is a black box and can include multiple layers in it.

The neuron/s at the hidden layer will take the input from the previous neuron and by using the weights of the axons (those connecting two neurons) and adds them up using a combination function. This is then inputted to a non-linear activation function like the sigmoid function to provide an output for the downstream neurons. This process is continued until the network arrives to the output layer and produces an output.

Since the input values have to be normalized, the output generated by an artificial network will have to be de-normalized to get the actual output values. What gave artificial neural networks the recognition and praise they have is their learning ability. The network learns by comparing the output value generated by the output layer with the actual output value of the target variable. Provided that there are enough records to feed to the network, for each instance or records the

output neuron have generated a prediction, the network will calculate the square of error and sum them. This is called the sum of squared errors or SSEs.

$$SSE = \sum_{instances} \sum_{ouptut\ neurons} (actual \quad output)^2 \qquad (1)$$

In the beginning stage, the weights used to connect the neurons between the input layer and hidden layer are chosen randomly. After the hidden layer, sigmoid activation function value of the previous neuron is used to estimate the weight. Once the feed forward part is completed and an output produced, the SSE can be calculated. The major objective of back-propagation now is to find weight values for these instances that will minimize the SSE. To do that, it uses an optimization technique known as Gradient Descent, which states "the gradient of SSE with respect to the vector of weights, w, is the vector of partial derivatives of SSE with respect of each of the weights." (T.Laroese & D.Larose, 2014).

The rule for back propagation is then given by:

$$w_{ij,\ \text{new}} = w_{ij,\ \text{current}} + \Delta w_{ij} \qquad \text{where} \quad \Delta w_{ij} = \eta \delta_j x_{ij} \qquad (2)$$

Where:

- w_{ij} is the weight for the it input to node j;
- η represents the learning rate choosing either too small or too large values for learning rate will make the network either too slow to find optimal solution or even unable to reach the optimal value;
- x_{ij} signifies the it input to node j and
- δ_j represents the *responsibility* for a particular error belonging to node j. The error responsibility is computed using the partial derivative of the sigmoid function with respect to net j (T.Laroese & D.Larose, 2014).

2.2.3.2.2. APPLICATION IN PRELIMINARY COST ESTIMATION

Siqueira (1998) developed a cost estimating model using Neural Networks for low-rise prefabricated structural steel buildings. He used data from 75 completed building projects for training, testing and evaluating the models with a 60%, 20%, 20% division. The accuracy of the model was very good as it has a mean absolute error of 6.56%.

Emsley et al. (2002) used a database of over 300 building projects to train neural network cost models, and the data obtained included final account balances. With a MAPE of 16.6 percent, the model was able to estimate the total cost of construction projects at the early stage.They also developed regression models to use as a benchmark for comparison and concluded that the use of ANN over linear regression is that artificial neural networks are more capable in modeling non-linearity in the data and, as a result, have lesser MAPE as compared to the most accurate regression model which had a MAPE of 27.9%.

Gunaydn and Dogan (2004) created an ANN model to estimate the cost of a square meter of building structural system in the early stages of design procedures. They gathered cost and design information from 30 projects ranging from 4 to 8 stories in Turkey. The trained ANN model's input layer consisted of eight parameters that were available during the early design stage. When tested, the trained ANN model had an accuracy of 93 percent. Similar results and conclusion were also reached by other researchers and is summarized in table 2.

Coming to Ethiopia, Tadesse and Dinku (2017) used cost data from 48 projects and developed ANN models having a MAPE of 32.58%. Though the data from these projects was converted to the base year value, which cost indexes they used was not presented. No published research was found with regard to building projects.

Table 2: Performance of ANN in Conceptual Cost Estimation of Building Projects - A Literature Summary

Author/s/year	Title	Summary and conclusion
Yadav and Vyas (2016)	Development of cost estimating method for bricklayer cost	• Used 23 years of data and developed ANN models to forecast Structural cost of residential buildings.
Dagbui and Smith (2012)	Neural Networks for Modelling the Final Target Cost of Water Projects	• Used data from 98 water related projects in Scotland and developed different models. • Aim was to estimate the normalize target cost

		• Concluded that ANN captures the relation between the target variable and attributes very well.
Arafa and Alqedra (2011)	Early-Stage Cost Estimation of Building Construction Projects using Artificial Neural networks	• ANN models were used to estimate the cost of building projects in the Gaza Strip. • 71 building projects' data was used • The ground floor area and number of storeys were among the most influencing parameters. • Similar conclusions was reached by …. Who used 169 building data from the Gaza strip to estimate the cost of building project
Emad et al. (2014)	Conceptual Cost Estimate of Libyan Highway Projects Using Artificial Neural Network	• To predict the conceptual cost of highway construction projects in Libya, data from 67 already built projects was used. • The model presented a minimum average percentage error (MAPE) of 2.86%. • Presented a methodology to account for inflation.
N.Bhirud and Vinayak (2017)	Pre-Design Stage Construction Cost Prediction of Building Projects Using Artificial Neural Network	• Used 12 building projects data and developed ANN model to estimate the cost of building project at the pre-design stage. • Instead of producing just the final cost, the model was used to estimate 18 different works involved in the project.

		• Evaluated the model in 3 actual projects. • Model showed an accuracy of 86.11% when estimating the total cost and an average accuracy of 75.55% when estimating the different work types.

2.2.3.3. FUZZY LOGIC

2.2.3.3.1. LOGIC

Let's assume a site engineer described the weather as 'hot' in a site diary. 'Hot' by itself is a fuzzy or vague way of describing the state of the weather. While us humans can determine and understand these vague descriptions of a parameter or state, computers generally are not good at understanding them. Computers understand values, particularly numerical and Boolean values like 0 and 1 where 0 represent a state of absolute falseness and 1 represent absolute truth of an instance.

If one is to use the weather parameter set out above as an input in a Neural Network model, for example, it is necessary to change these vague linguistic, qualitative value into quantitative Boolean values by using rules. But the Boolean domain is a set that only includes 2 elements and they are built on the premise of 'either…or' and cannot exactly represent the states in the middle range.

For instance, we can use 1 to describe hot weather and 0 to describe a cold weather but the problem with this is weather conditions described as warm, slightly cold/hot and so forth are not exactly represented by it. Techniques like multiple Regression, SVM/R and artificial neural networks are inherently quantitative and require the input parameters to be presented in a numeric format. Fuzzy linguistic parameters cannot be properly represented using a Boolean value and this means we will have to use different techniques to group or classify certain parameters that are better described qualitatively and convert them to a numerical value.

Fuzzy Logic or the logic of fuzziness is an ideal solution when we want to refer to things that are described in a vague or unclear manner. We can consider Fuzzy sets as special kind of sets

that allows for elements to be members of a set to a certain degree. Fuzzy sets can help us express or describe to what extent a statement is true or false; in the case of our example, to what extent the weather is hot or not. This is done through different fuzzy rules, often described as if-then statements and a function that identifies the degree of membership of an element in a set - i.e., the degree of hotness (the element) in the set "weather". This function is known as membership function, μ. The fuzzy inference system has three major processes in its core: Fuzzification, fuzzy logic operation and Defuzzification

Fuzzification is the process of transforming a quantitative (or crisp) value into a fuzzy verbal input. Once fuzzified, a controller comprised of fuzzy rules and a membership function performs the fuzzy logic operation and assigns the output. The output generated by the controller is a fuzzy one. In order to make sense of the output, defuzzification will be carried out; which will convert the fuzzy output to a crisp one (ElProCus Technologies, 2020).

2.2.3.3.2. APPLICATION

Sawalhi (2012) conducted research to assess the applicability of Fuzzy Logic in estimating the parametric cost of building construction projects in the Gaza strip. The research first identified 5 important factors that influences building costs and developed six membership functions and 27 rules. Sawalhi concluded that fuzzy logic model provided "a prediction near the actual cost of a test project". However, no numerical performance value was provided on the performance of the model.

2.2.3.4. SUPPORT VECTOR MACHINES

SVM (Support Vector Machine) is a supervised learning algorithm which is mainly used to classify data into different classes. Unlike most algorithms, SVM makes use of a hyper plane which acts like a decision boundary between the various classes. A hyper plane is a plane in n dimension that can help classify a data. If we are classifying a data into 2 classes, we will have a 2-dimension hyper plane, which is a line. In a 3-dimensional data, the line becomes a 2-dimensional plane and so forth. In case of data mining, the number of dimensions can be decided based on the amount of attribute or features we have.

SVMs work by minimizing the distance between the hyper plane and data points near to the hyper plane, known as support vectors. The closer the hyper plane can get to these support vectors, the better or more accurate will be at classifying among the data points (Gandhi, 2018).

SVM can be used for classifying non-linear data by using the kernel trick. The kernel trick means transforming data into another dimension that has a clear dividing margin between classes of data. After which you can easily draw a hyper plane between the various classes of data. Kind of like how humans solve problems by evaluating them from a different perspective, except that the perspective in this case is another dimension. The main advantage of SVM is that, the same algorithm can be used for both classification and prediction or regression problems with slight adjustments. When used in classification problems, it is called Support vector machines whereas, when used in prediction problems, it is called Support Vector Regression.

Support vector regression models have proved to be powerful in predicting cost estimates as shown by different literatures. After identifying and ranking the factors affecting costs of road projects based on data collected from 70 road projects in the Gaza strip, the researcher developed a prediction model using Support vector machines. The model developed had a MAPE of only 5% (El-Sawalhi, 2015).

Petruseva, et al. (2017) conducted comparative research to assess the accuracy of SVM and regression models. Using data from 75 structures built in Bosnia and Herzegovina, they developed SVM and regression models to predict cost of building projects. To make the comparison, they used similar predictor variables for both models. The results indicated that SVM or SVR in this case was much better at predicting cost of building projects when compared to the regression models (Petruseva et al., 2017).

2.2.3.5. DECISION TREES

A decision tree is a machine learning algorithm that recursively split the data into subsets, starting with a binary split and continuing until no further split can be made or a stopping criterion is met.

A decision tree building step can be summarized as follows: First, the complete training data set is saved in a root or parent node. The root node is then split into two nodes (called daughter nodes). These daughter nodes are associated to a particular subset of the training data, S, and are split based on the value, V, of specific attribute A such that;

$$S_{left} = \{s \in S : s(A) \geq V \;\; And, \; S_{right} = \{s \in S : s(A) < V \qquad (3)$$

Where; S_{left} and S_{right} indicate split for daughter node (Simone, 2012).

The value, *V*, is also called a splitting point, meaning, instances having an attribute with a value greater or equal to V are split to the left daughter node and the remaining are partitioned to the right daughter node. The daughter nodes are further split based on other values of attributes.

The splitting point is found by searching over all possible attributes and values and choosing the variable and split that reduces the impurity function. The impurity function is different for classification and regression problems. For classification trees, the impurity function takes a form of the Gini Index or the Shannon entropy whereas for the regression trees, the mean square error is used. Accordingly, the splitting attribute *j* and splitting value (point) s is one that solves:

$$min_{j,s}[min_{c1}\sum_{x_i\in R_i(j,s)}(y_i \quad c_1)^2 + min_{c2}\sum_{x_i\in R_2(j,s)}(y_i \quad c_2)^2] \qquad (4)$$

Where the criterion for minimization,

$$c_1 = ave\left(\frac{y_i}{x_i}\in\ R_1(j,s)\right) And\ c_2 = ave\left(\frac{y_i}{x_i}\in\ R_2(j,s)\right) \qquad (5)$$

And the regions of partition, $R_i(j,s)$ are given by:

$$R_1(j,s) = \{X|X_J \le S\}\ and\ R_2(j,s) > \{X|X_J S\}\ (Trevor\ Hastie,\ 2017) \qquad (6)$$

This process for the splitting is done in a recurssive manner until the tree cannot be further partitioned either due to lack of data or as a result of some stopping criteria stated. The final node is called the terminal or leaf node. In prediction of a continous variable, the terminal node value is calculated by calculating the mean of the values of the data points left within the node.

The size of a decision tree describes the complexity of the model. Ideally, we would want a tree that is not too big but still can generalize well. As discussed above, the algorithm will not stop splitting untill there is no data left or we provide a stopping rule for it. The need for this stopping rule is that as the tree keeps splitting more and more, the fewer number of cases will be left in the terminal node. This will create a high varaince error during prediction, thereby affecting it's performance. Furthermore, by fitting the training data too 'perfectly' (overfitting), its' performance on an unseen data will be greatly compromised.

Thus, to avoid overfitting, certain rules must be put in place so that the tree size is reduced by cutting or prunning some of the leaf nodes and turning the branch nodes into leaf nodes. This process is called prunning and there are two different prunning methods: Post and Pre prunning. During post prunning, the decision tree is first built and then some of the leaf nodes are cut based on their information gain (or the power they have in increasing the accuracy of the model). The rule or critera in this case is the minimum information gain a node shall have. Accordingly, those nodes that donot meet the minimum information gain are cut from the tree. Whereas in the case of pre-pruning these nodes are not built in the first place.

Decision Trees are among the most popular machine learning algorithms used in different areas. Their interpretablity has given them advantages in most classification problem. Decision trees have an added advantage of handling non linear data when compared with linear regression (Shin, 2015) developed regression trees for perliminary cost estimation of building projects using 234 buildings data, The regression tree model was found to have slight improvement in performance when compared with Linear regression,neural network and Support vector regression.

2.2.3.6. GRADIENT BOOSTED TREES

A gradient boosted tree (GBT) is an ensemble model that uses decision trees as a base learner to develop a more complex model by using boosting techniques. Gradient boosting is based on 3 elements. These are: optimization of loss function, a weak-learner and an an additive model that adds those weak learners and minimize the loss function. The objective of any supervised learning algorithm is to define a loss function and minimize it. The loss function to be used depends on the task at hand, Regression models can use squared error whereas classification tasks could use logistic logarithmic loss. As long as it is differentiable, the algorithm accepts different loss functions.

Regarding the weak learners, Decision trees are mostly used as the weak learner in gradient boosting although, regression has also been used. The decision trees are constructed in a greedy manner, choosing the best split points based on purity scores like Gini or to minimize the loss.

Gradient boosting requires the weak learners to remain weak. In order to ensure that, size of the decision trees could be restricted using preprunning. These weak Trees are then added one

at a time while existing trees in the model remain unchanged. A gradient descent procedure is used to minimize the loss when adding trees. algorithm follows this process:

As an input, the algorithm requires a dataset $\{(x_i, y_i)]\ w\ \ ere\ i = 1, \dots n$ and a differentiable loss function $L(y, F(x))$ where $F(x)$ is a function that gives us the predicted values.

Step I. Intitialize model with constant value. This means to predict an initial value for the model, by finding a predicted value that minimizes the loss function, i.e.

$$F_0(x) = argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma) \qquad (7)$$

where γ is the predicted value for the intialized model.

Step II: for number of trees, m=1 to M:

A: Compute the pseudo residual which is the derivative of the loss function with respect to the predicted value, F(x).

The pseudo residual is given by:

$$r_{im} = \left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \qquad (8)$$

$for\ i = 1, \dots n\ w\ \ ere\ i\ is\ t\ \ e\ sample\ number$

B: After the residual is calculated for each sample, a regression tree is built to predict the residuals rather than the weights. The leaves are then become the terminal nodes $R_{jm}\ for\ j = 1 \dots . Jm$

C: *For J = 1…. Jm, Compute* $\gamma_{jm} = argmin_\gamma \sum_{x_i \epsilon R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$ *(9)*

For each leaf in the new tree, calculate an output value of the residulal γ_{jm} that minimizes the loss function $L(y_i, F_{m-1}(x_i) + \gamma$ for leaves 1 to Jm.

D: The prediction for each sample is then updated using equation 10. The prediction of the second tree is then the sum of the prediciton of the first tree plus learning rate times residiual prediction made in step C for the residual output value. This process is then iterated M times.

$$F_m(x) = F_{m-1}(x) + \mu \sum_{j=1}^{J_m} \gamma_m I(x \epsilon R_{jm}) \quad (10)$$

Step III: after iterating M times, Output the last tree result, $F_M(x)$ as a final prediction.

2.2.3.6.1. APPLICATION IN COST ESTIMATION

Elmousalami (2019) in a research where he compared 20 machine algorithms to estimate the preliminary cost for Field canal improvement projects, has adopted a stochastic gradient boosting and another variants of the gradient boosted tree algorithm called Extreme Gradient boosting (XGBoost). The result indicate that while the XGBoost was superrior of all 20 models, the Stochastic Gradient boost didn't perform as well as a Multilayer perceprton.

Chakraborty et al. (2020) developed cost predction models using LR, ANN, Random Forest and Gradient boosted trees. They found out that the Grardient boosted tree model had the least error with a RMSE of 0.5 while the LR, ANN and Random Forest models had RMSE of 1.8, 0.9 and 0.7 respectively.

2.2.4. A COMPARISON BETWEEN DIFFERENT TECHNIQUES

2.2.4.1. LINEAR REGRESSION VS ARTIFICIAL NEURAL NETWORKS

Different researches have also been carried out to compare the accuracy of regression and AI methods in the cost estimation of building projects. Sonmez (2004) used cost data from 30 retirement community projects in the USA to develop cost estimation models using regression and ANN techniques. He used the 30 projects to develop the models. The selected regression model and the two NN models were compared using the mean squared error (MSE) and the mean absolute percentage error (MAPE). The results indicated that NN models provided better fit to the data (using all 30 projects) than the regression model; However, the regression model had a better prediction performance than the NN models.

Their result was negated by Cho et al. (2013), who also compared a regression model and a NN model to estimate the cost of elementary school construction. The regression model was developed using SPSS and the cost data from 96 schools, of which 76 were used for model development and 20 for testing. NeuroSolutions (2012) - an Excel based ANN modeling software was used to build the NN model. The accuracy of the models was tested by comparing

the MAPE and it was found that the error rate of the NN model was lower than that of the multiple regression model.

2.2.4.2. LINEAR REGRESSION VS SUPPORT VECTOR REGRESSION

Considering the higher generalizing capability of SVM, it is expected that Support vector machines will be superior to linear regression in estimating conceptual costs. Similar conclusion was also reached by Petruseva et al. (2017) who compare the performance of these two methods in estimating cost for building projects at the conceptual stage. They based their models on 75 constructed buildings data and determined that SVMs have a much better accuracy (MAPE of 0.3) than Linear Regression models (MAPE 4.79%).

2.2.4.3. ARTIFICIAL NEURAL NETWORK VS. SUPPORT VECTOR REGRESSION

Peško, et al.(2017) used 166 projects data to estimate the cost and duration of road projects. As the objective was to compare the accuracy of SVMs and ANNS in estimating the cost and duration of road projects, they developed a number of models to choose the best from. Accordingly, SVM was found to outperform ANN in estimating the cost as the MAPE of the most accurate SVM model was 7.06% whereas the most accurate model of ANN only provided an average percentage accuracy of 73.74%. (Peško et al., 2017)

Patil and Salunkhe (2020) on the other had used 147 data from projects in India and built cost estimation models using 3 different techniques namely, Regression, Support vector Machine and Artificial Neural Networks. The author concluded that while all 3 methods provided a good enough estimate, the ANN model was the most accurate one with a MAPE of 0.15%. SVM came in second with a MAPE of 0.3% and the regression model had a MAPE of 4.79. This result is in conflict with the study of Patil and Salunkhe (2020).

Similar result was also reached by Gwang-Hee et al. (2013) after comparing ANN, SVM and regression in estimating construction costs for building projects. The regression models were developed using the stepwise techniques in SPSS (a statistical software) and the construction costs of 217 school building projects in Korea. The performance of the models was measured using the MAPEs. ANOVA was used to test the null hypothesis that the MAPEs of the three approaches were equal. They concluded that NN model, in this case, showed more accurate estimation results than the RA and SVM models (Gwang-Hee et al.,2013).

2.2.5. A HYBRID APPROACH.

Though using a single machine technique alone like, NNs or Fuzzy Logic, Case based reasoning, etc. has shown to give satisfactory results in estimating preliminary costs, Cheng et al. (2009) argued that these approaches have their own drawbacks. In order to solve this and achieve an even better predictive accuracy, different researches have been done by trying to fuse two or more of these techniques into a single, more powerful one.

2.2.5.1. GENETIC ALGORITHM AND THE POWER OF OPTIMIZATION

"it is not the most intellectual of the species that survives; it is not the strongest that survives; but the species that survives is the one that is able best to adapt and adjust to the changing environment in which it finds itself". – ***Charles Darwin, Origin of Species***

Genetic Algorithm (GA) is an algorithm developed to solve both constrained and unconstrained optimization problems. Genetic algorithm belongs to a group of Evolutionary algorithms which are based on the theory of Evolution or natural selection. Genetic algorithms work by creating a set of possible number of solutions to one problem, which are called population and by allowing these population to go through the process of evolution. The general steps involved in GA is provided in Figure 3.

These individual solutions in the population are given fitness scores which dictate their ability to compete amongst other individuals in the set to become the optimal solution for the problem. Accordingly, Individuals with higher fitness scores are then selected as they are genetically predisposed to become optimal solution than those with lower fitness scores. The selected individuals are then grouped in pair and reproduce, as part of the evolution cycle. This allows the set or population to have even more possibly good solutions for the problem. This is called Crossover (Mitchell, 1998).

When individual of similar traits keeps reproducing, diversity dies, and along with it, the population. In the real world, natural organisms will have to evolve to adopt to their current situation and survive, the process of natural selection will choose which kind of mutation are beneficial and keep that while those mutation not beneficial become less common over time and die out. In Genetic Algorithm, mutation helps by increasing the diversity in the population so that the algorithm will not converge to the local optima solution. This process will continue until the optimal solution is reached by the system; thereby optimizing the system at hand.

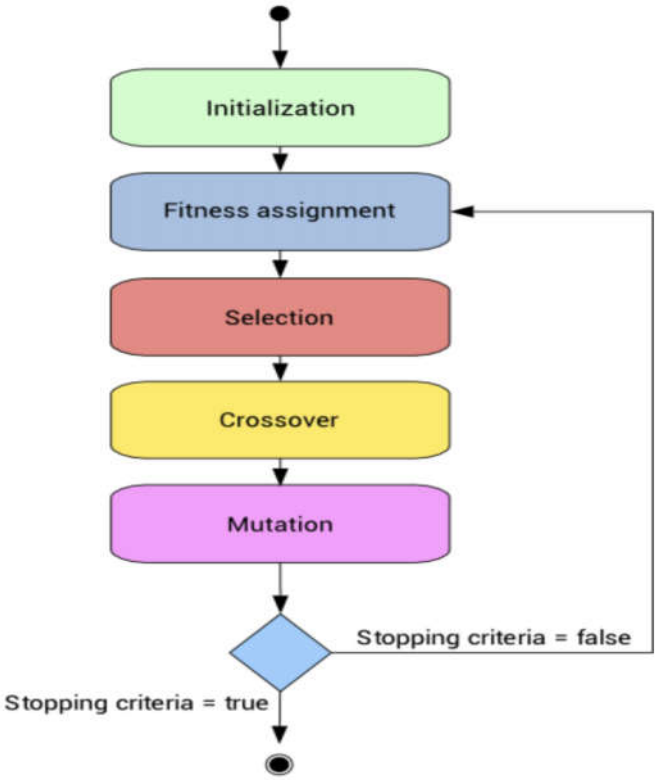

Figure 3: General Procedure of Genetic Algorithm

2.2.5.1.1. GENETIC ALGORITHM AND NEURAL NETWORK

Genetic Algorithms can be applied to neural networks in different ways. The first method was identified by David Montana and Lawrence Davis, who used GA instead of Back propagation to determine the weights for the connection between the neurons. This was done to see if GA can help the network not get stuck in the local optima. They concluded that GA can outperform BP in some cases (Montana & Davis, 1989).

A compromise, perhaps, is to use genetic algorithm to optimize BP instead of using a one or nothing approach. This system was adopted to model construction cost estimates by Guangli Feng (2013) and provided "a faster and effective result". The second method involves optimizing the number of hidden layers required to get the most accurate NN model using GA instead of using guesswork and trial and error. The third method is to use GA to identify the learning rate for the NN model (Mitchell, 1998).

Kim et al. (2004) used a combination of second and third methods to develop a neural network model to estimate the construction cost of residential building projects in Korea. He based his

research on 530 historical cost data collected in a span of 3 years. To determine the performance of GA in identifying the ideal number of hidden layers, he developed another model using the trial-and-error method of choosing hidden layers. The most accurate model using GA optimization has a minimum average error rate of 2.62%. Which is quite good. The research also noted that the either too few or too many hidden layers will have an impact on the accuracy of the model as models with little hidden layers have lesser learning capacity and models with too many hidden layers will cause a problem of overfitting – a problem that occurs as a result of having too few data to train on while having too much processing capacity. Overfitting will also take too much time to process (Kim, et al., 2004). In all these methods, GA was used to optimize the process of developing neural networks thereby making it faster to arrive at more accurate results.

2.2.5.1.2. CASE BASED REASONING AND GENETIC ALGORITHM

The problem with standard CBR approach as noted by Ji et al. (2011) is that these methods lack the ability to measure the similarity among cases when the target case is outside the case base range set out. To address this issue, they adopted Genetic Algorithms as a way to measure similarities. The check the performance, they deployed two CBR models that estimate the required budget of apartment buildings in Korea – one based on GA and the other based on inductive method. The results of the research showed that the genetic algorithm based CBR model provided more accurate results.

Jung also used GA and CBR to estimate the preliminary cost of different projects like office buildings, facilities and public apartments. Important predictors were chosen and ranked using correlation analysis for each project cases/types local search methods were used to identify the weights of each predictor/attribute instead of using random weights for the GA. The models developed provided better performance as compared to the ones modeled only using CBR (Jung, et al., 2020).

2.2.5.2. NEURO-FUZZY MODELS

The inherent drawback of Neural Networks is that it is difficult for users to understand the logic the networks are using to determine the output, another thing Neural Network lack is the ability to incorporate vague, complex qualitative information in its system. A Fuzzy system on the other hand provides users with a way to learn from these fuzzy inputs and provide an interpretable result in a form of rules but it lacks the incredible learning capability of Neural Networks. Ergo, the birth of the hybrid Neuro-fuzzy systems. Neuro-fuzzy systems can be further classified as co-operative and concurrent neuro fuzzy system. While in the former case, the output of a fuzzy system is fed as an input to the neural network system and vice versa, in the latter case, the two systems work together as one (Wu, Zhang, & Du, 2011). Neuro-fuzzy systems have been adopted in different industries for different purposes like decision making, forecasting, simulation and optimization (Getaneh & Sumati, 2020) .

Wang, et al. (2017) used neuro fuzzy systems known as FALCON (fuzzy adaptive learning control network) to estimate the conceptual cost of a building project. It based the model on 46 historical building data and used regression and multi-factor evaluation model to evaluate and identify project characteristics with highest impact on the project cost of buildings. The proposed model had an average accuracy of 91.82% when tested on 3 different projects, with the minimum and maximum accuracy being 89.82% and 92.96% (Wang et al., 2017).

In relation to cost forecasting, Yu (2006) used a quantity and unit price estimation Model he named the PIREM model and the adaptive Neuro-fuzzy system (ANFIS) to estimate the cost of residential projects in china. The result showed that neuro-fuzzy system was more than capable of predicting costs for residential projects in the conceptual phase.

2.2.5.3. EVOLUTIONARY FUZZY NEURAL INFERENCE MODEL (EFNIM)

Cheng et al. (2008) propose combining Fuzzy logic, ANN and Genetic algorithm together which, in the process, will eliminate the individual drawbacks and maximizes the strength of these individual approaches. What they came up with is a technique called, Evolutionary Fuzzy Neural Inference Model (EFNIM).

The EFNIM combines Genetic algorithm, Fuzzy Logic and Neural networks into a single approach that can be used to estimate the costs of construction projects, particularly building projects. Though the use of Fuzzy logic and neural network is not a new thing, particularly due

to the added advantage of using NNs to solve for Fuzzy logics drawback in self-learning. On the other hand, fuzzy logic helps the model by helping it understand vague data through heuristics and taking these fuzzy sets as weights for the neural networks. The other drawback of fuzzy logic is with regards to identifying the appropriate membership function distribution, which is used to identify the degree of truth in fuzzy logic. This is where Genetic algorithm's ability to optimize comes in and identifies the optimal membership function.

Using this model and cost data from 28 building projects in Taiwan, the researchers developed to cost models namely; total construction cost model and category cost estimation model, which estimates cost for different categories/phases of the construction projects like, geotechnical, electromechanical, interior decorating, structural. Of the 28 projects, 23 were used as training data sets and the remaining 5 were used for testing (Cheng, et al., 2008).

The result showed the model was able to estimate the total cost and categorical costs with an average error of 16% and 6.7% respectively. Though the researchers implied this model provides a superior result than that of the individuals, they have not base this with their own results, as a result, it is any body's guess if or by how much EFNIM is superior to Fuzzy logic or neural networks in estimating construction costs.

Cheng, et al. (2009) added up on Cheng et al. (2008)'s work and developed an Evolutionary Fuzzy Hybrid Neural Network model. This model incorporates the EFNIM model with an addition of a higher order neural networks instead of the linear Neural Networks. Higher order neural networks are made of combination of higher order polynomials as the inputs and weights (Gupta & Bukovsky, 2012). This fusion of neural networks is known as Hybrid neural networks (HNN). The authors also compared the performance of EFNIM and EFHNN and concluded the evolutionary fuzzy hybrid neural network model showed better performance in predicting conceptual costs and credited the higher performance to the HNN models. Dalhoum & Al-Rawi's (2019) theorem seem to reject that notion and shows Higher order neural networks and ordinary neural networks delivers equivalent performance and whatever slight increase in performance that maybe gained from using higher order neural networks is not worth the hassle of developing them.

2.3. SECTION III - QUANTITY-BASED APPROACH

All these statistical and ML techniques discussed previously have one thing in common – They use historical cost data to predict or forecast costs for future projects. Although different researches have shown that these methods can be used to estimate the potential cost of projects at the preliminary stages, it shall be noted that as they are based on cost data of previously built projects, inflation will need to be taken into account in the estimate (García de Soto, 2014).

In Table 3 a description of a number of studies and the methods they adopted to take inflation into account is presented. While some ignored this inflation into account, one approach that most researchers have taken is to adjust the cost data for inflation based on particular inflation index variables or price indexes of the region. Other researchers like Peško et al. (2017), Cheng et al. (2009) and Jung et al. (2020), however, did not discuss if or how they accounted for inflation in their models.

In countries where market is stable and material and labor prices don't fluctuate unreasonably, the construction costs may reasonably be adjusted for the increase or decrease based on cost Indices. Even then, these countrywide price indexes are not representative of specific industries, thus, will not provide most accurate results. A report by the Canadian construction association (2012) identified using non-industry specific cost indices as among the reasons for poor performance of estimation models in the construction industry – a notion supported by Zhang (2017), who, after accounting for economic fluctuations as dependent variables in his models, concluded that these models will not work in environments with high economic fluctuations.

Unfortunately, In Ethiopia's case, the market is far from a stable one. Due to reasons like shortage of currency and other Political and socio-economic problems, the cost for materials fluctuates quite rapidly - mostly increasing. According to research by Asteway (2008), price fluctuation in the country occurs in an unpredictable manner with increase in material prices by more than 34%.

In his thesis regarding the causes of price escalation and the effects it has on price adjustments in Ethiopian federal road project, Mohammed (2013) identified that price fluctuation as a major cause for problems regarding inaccurate inflation rates and construction price indices in Ethiopia's Federal Road projects. This issue will make it difficult to estimate the cost for the

building projects based on past cost data. Luckily, researchers have also devised another solution to this problem: Quantity-Based Preliminary Cost Estimation.

The logic behind this method as noted by García de Soto (2014) is that, if we could reasonably estimate the cost of a project using these techniques, we can then use the same techniques to estimate the quantity for different work items of the projects and then use the estimated quantities and a current price rate for the work items to estimate the final cost of the project. This ensures a clear separation between technical estimates (e.g., quantities) and market fluctuations (e.g., cost of materials and labor) during the early stages of a project. This also allows managers to make better decisions and keep a better track of their projects by controlling the changes in cost and quantities separately (Borja & Bryan, 2016).

Table 3: Methods Used to Account for Inflation in Different Researches

Author	Objective	Accounts for inflation
Emad and El-Fitory (2014)	To estimate Conceptual cost for highway projects in Libya	Used Inflation rates of the country But model is open for user to input specific inflation rates.
Wang et al. (2017)	To estimate Conceptual cost for building projects	Costs were adjusted for inflation but does not describe which inflation rates are used.
Gardner (2015)	To estimate cost of highway projects at the conceptual stage	Used a constant nominal inflation rate of 3% based on historical averages Recommended further research should be carried out to identify a method to predict a more accurate an inflation rate
Alemayehu (2014)	To estimate cost of road projects in Ethiopia	Used cost indices from CSA to adjust for inflation
Tadesse and Dinku (2017)	To estimate highway project costs in Ethiopia	Used Consumer Price index (CPI) from the Central statistics agency of FDRE to adjust for inflation
El-Sawalhi (2015)	To estimate the parametric building construction project cost	No adjustment for Inflation has been discussed

Lesniak and Zima (2018)	To calculate cost of Construction projects using sustainability factors	Unit prices were adjusted for Inflation based on price indexes from a news paper
Patil Salunkhe (2020)	To compare these methods in estimating cost of a project	No adjustment for Inflation has been discussed

2.3.1. PERFORMANCE OF QUANTITY-BASED MODELS

A number of researches have been carried out to asses' quantity-based estimation models since the 1990s. Yeh (1998), one of the earliest resercher on this area, developed a neural network (NN) model to estimate the required quantities for steel and reinforced concrete (RC) buildings, using data from 400 buildings (300 for training and 100 for testing), and compared the performance of the model with that of cost-based models developed using NN, linear regression, and nonlinear regression. He concluded that use of quantity based NN models produced more accurate result than the other cost-based models.

Singh (1990) developed a computer-based cost model to estimate the cost of reinforced concrete beam and slab construction in high-rise commercial buildings based on the quantity of works. He used Completed projects to compare the actual quantities with the ones obtained using the computer model. The comparison indicated that actual quantities used in different projects were always more than those calculated by the computer model. The difference varied from project to project, with an overall range of 5-17%. The computer model was calibrated to compensate for that difference. He compared the results with some local projects and concluded that the computer-based cost model provided practical value to the industry as it could give the desired information quickly and accurately.

Bakhoum et al. (1998) did a study regarding the estimation of quantities of concrete, reinforcement and pre-stressing required in pre-stressed concrete bridges over the Nile River in Egypt. Then, a unit price could be assigned to the estimated materials to determine the cost of those structures. A preliminary quantity estimate for concrete approaches and navigable spans was conducted based historical data and data of bridges being built during the time of the study, 10 bridges were used for that analysis. A total of 6 variables were finally considered

as the ones having the most influence on navigable spans superstructure during the preliminary design.

In addition, using the BrainMaker simulator by California Scientific Software, Inc., NN models were developed using 5 input variables and data from 22 bridges to train and test the NN. Although the study does not provide information about the results from the models or their performance, the authors recognize the benefits of quantities estimates to contractors and owner, depending on the contract type (Bakhoum et al., 1998).

Recently, Son et al. (2013) used the wall to floor ratio as the only predictor variable to predict quantities for concrete, reinforcement, and formwork in basement, ground, and upper floors. They concluded that their model had a better performance than the cost-based methods with errors ranging from *-4.73% to 4.34%*. Furthermore, they outlined that their model was responsive to market and design changes. So, by changing design variables such as floor height, floor use etc. on the model, the corresponding changes in the cost can be analyzed, hence making it easier to test different design alternatives to meet budget.

García de Soto (2014) proposed a Methodology using Artificial Neural networks for developing preliminary construction material quantity (CMQ) estimate models and applied his methodology in estimating the quantities for construction of tall frame structures in cement plants. The models provided the quantity estimates with an error ranging from -13% to 17% with 72% of the estimated quantities having a percentage error below 5%. He concluded that as these models were very much capable in estimating the quantity of tall-frame structures for future cement plants at the preliminary stage.

Dursun and Stoy (2016) adopted a novel approach called multi step ahead (MSA) to estimate cost of building projects in Germany. In this approach, the quantity of works is first estimated using ANN and Regression first. It then uses these estimates as predictor variables to the cost model. This allows the cost model to be modeled on more information than what it would normally be modeled with in preliminary stages. After doing a comparison, it was concluded that the MSA approach outperformed the conventional cost-based approach (Dursun & Stoy, 2016).

Oluwafunmibi and Lam (2020) used Support vector regression to develop models for predicting the quantities of reinforced concrete structural elements. They developed 12 models

in total and used parameters like Gross Floor loading, gross floor area, Live load, building footprint, height and number of floors as predictors for these models.. The authors also used bootstrapping; a technique used to improve the accuracy in models when little data is involved. The bootstrapped models were able predict the quantities of works within accuracy interval of 95% (Oluwafunmibi & Lam, 2020).

Aside from building projects, other researches have also been carried out to estimate the quantity of works for different kinds of Civil Works. A quick summary of these researches is provided in Table 4.

Table 4: Quantity Based Approach Adopted in Other Civil Works

Author/date	Estimated	Technique used	Objective
Chou et al. (2006)	Highway Repair Project	Regression analysis	Estimated quantities of different works involved in highway repair
Kim et al. (2009)	Pre-stressed Concrete bridges	Regression Analysis	Estimated standard work quantities involved in pre-stressed concrete bridge construction
Park et al. (2013)	Substructure for steel box girder bridges	Regression analysis	Estimated quantity of materials needed to build substructure part of steel box girder bridges
Du and Bormann (2014)	Power Plants	Case Based Reasoning	Estimated quantity of works involved in power plant construction

Though these methods have shown higher accuracy in conceptual cost estimation, the problem with this approach as noted by Borja and Bryan (2016) is it becomes too cumbersome as different models will need to be prepared for the different works that make up the buildings. Although, this could be solved by using models that allow Multiple input and Multiple Outputs, there has been no published research on it.

2.4. SECTION IV - IDENTIFICATION OF FACTORS AFFECTING PREDICTION VARIABLE

Any predictive model bases its prediction of the dependent variable, also called the prediction, on a number of independent variables, also known as features or attributes. The performance of the predictive model is highly dependent on the factors it uses to estimate or generalize on. Thus, identification of those factors that have the most impact in the cost of a project (or the quantity of work being estimated) is an utmost importance. Identifying them alone, however, is not enough. Many Machine learning algorithms require a certain balance between the number of predictor variables and the number of instances we have in a data set. If the number of instances is too few while we have too many features, these algorithms tend to underperform as there will be too few data to learn from and associate these features with the dependent variable.

These factors will therefore need to be ranked using some metric so that only the most important factors can be chosen and used in the model. This is called dimensionality reduction and it is a crucial step needed to be taken particularly in small data sets. There are different methods that can be used to identify and rank these features. These methods can be broadly classified into qualitative and quantitative methods.

The qualitative method entails doing questionnaires and interviews with experts in order to identify these cost driving factors. These experts' opinions are then evaluated by techniques like the traditional Delphi Method and Likert Scale and fuzzy analytical hierarchy process (AHP) to identify the most important attributes to use in the prediction model.

The quantitative method on the other hand is somewhat a tailored approach that is specific to the collected data. In this method, different statistical and Machine learning algorithms are used to identify and extract the necessary information from the data on hand. There are a number of algorithms that can be used to identify features with the highest impact on the prediction, of which Factor analysis and regression methods are the most common ones.

The simplest quantitative method is the correlation method. In this method, the relation among all features is displayed in a correlation matrix and those variables having correlation coefficients greater than 80-90% and lower than 30% are removed from the pool. This effectively removes those variables that are redundant and unimportant.

The second approach is Principal Component Analysis (PCA). While the correlation method focuses on removing variables, PCA focuses on combining these variables and creating a new variable that summarizes the given data well. This new variable is called Principal Component and it is not correlated with other variables (Elmousalami, 2019). As the original attributes are merged to create new attributes, the relation between individual independent variables and the output variable is not understood well, especially when complicated mathematical operations are used to combine these independent variables.

The regression methods include three approaches: Forward, Backward and Stepwise-selection. These approaches use linear regression to predict the output variable based on variations of the independent variable. What makes one approach different from the other is the method it uses to choose the independent variable. The forward selection begins the estimation process with zero attributes and adds one attribute at a time, evaluates the performance of the model. If the model has made significant improvement as a result of the addition of the variable, the predictor variable is allowed to stay. This process is done for each predictor.

The backward selection on the other hand begins with all attributes in the model and through tests like the t-test, the significance of the attributes is evaluated. The ones who affected the model's accuracy the least are eliminated and the process restarts until a specified minimum number of attributes is met.

The Stepwise-selection is a combination of forward and backward selection where the variables are both included or removed in each step. Genetic algorithm is also known to be used in feature selection by reducing irrelevant or redundant variables from models, particularly in ANN. Elmousalami (2019) surveyed 37 literatures to identify which feature selection method is used and concluded that majority of researchers adopted the quantitative approach whereas only 5 adopted the quantitative approaches, particularly correlation matrix, regression and Genetic Algorithm.

2.4.1. ATTRIBUTE/VARIABLE IDENTIFICATION

As discussed earlier, there are two general data mining types: Descriptive and Predictive. Predictive models are generally built on two types of variables: Predictor variables, also called attributes and the prediction variable, the output.

Aside from the general characteristics of the building, very few design information is available at the preliminary stages. Thus, the predictor values to be used in developing these estimates must be the characteristics of the buildings known at the preliminary or conceptual phase of the project. A literature review was carried out in order to identify these characteristics. Accordingly, different attributes that affect the final and structural cost as well as quantity for structural works of a building were identified from literature review are presented in Table 5.

Table 5: Influencing Attributes Identified from Literature Review

	Factor	**Authors and Remark**
1	Number of Floors	**Parameter type**: Numerical; **Authors:** Dursun and Stoy (2016); Wen-der and Skibniewski (2010); El-Sawalhi (2015); Sawalhi (2012); Pal et al. (2015) ; Bhirud (2017); Idowu (2019) **Used in:** Final Cost, Structural Cost and Quantity estimation models
2	Total slab area	**Parameter type**: Numerical; **Authors**: Wen-der & Skibniewski (2010); Bhirud & Ambrule (2017); Dursun and Stoy (2016) **Used in:** Final Cost, Structural Cost and Quantity estimation models
3	Foundation type	**Parameter type**: Categorical (Pile, Isolated, Mat etc.) **Authors:** Wen-der and Skibniewski (2010); Sawalhi (2012); Pal et al. (2015) **Used in:** Final cost and Structural Quantity estimation Models
4	Type of Slab	**Parameter Type**: Categorical; **Author:** Shehatto (2013) **Used in:** Final cost Structural cost and reinforcement quantity estimation models **Justification to use in structural quantity models:** According to Kiran and Issac (2018) and Belay (2004), the type of slab, particularly slab choice between ribbed and solid slab has an effect on the quantity of concrete work, Rebar and Formwork.

5	Slab Thickness (average)	**Parameter type:** Numerical **Authors:** Mahamid (2016) identified the volume of slab in cubic meters, which is a function of the slab area and thickness, has shown to have a direct (linear) relation with the reinforcement steel quantity estimation. He also identified the average slab thickness is related with the slab type of the building. **Used in** reinforcement quantity prediction models
6	Internal decoration	**Parameter Type:** Categorical (Basic, Standard, Luxurious) **Authors:** Pal et al. (2015); Wang et al. (2017) **Used in:** Final cost estimation models
7	Flooring Material type	**Parameter Type**: Categorical; **Authors**: Pal et al. (2015) **Parameter Type**: Categorical; **Used in**: Final cost estimation models
8	Duration for Project	**Parameter Type**: Numerical; **Authors**: Kim et al. (2004) **Used in**: Final Cost estimation Models
9	Shoring work	**Parameter type:** Boolean (1 for yes, 0 for No) **Author:** Hailemariam et al. (2020) identified Shoring work can cost as much as 5% of the total building Cost, making them critical in determining the final cost estimation. According to them, while the need for shoring work is primarily governed by geotechnical investingations, the site condition such as availability of ground water in the area,proximity to nearby buildings and the size and depth of the proposed building are among the main critereas that influence the need for a shoring work. As these critereas are known in the prelimiinary stage, shoring work is included in as one parameter.
10	Gross Floor Area	**Parameter type**: Numerical; **Author**: Kim et al. (2004); Jung et al. (2020) & Idowu (2019) **Used in:** Final Cost, structural cost and quantity estimation models
11	Total no. of Lifts	**Parameter Type**: Numerical; **Authors**: El-Sawalhi (2015); Sawalhi (2012); Bhirud and Ambrule (2017); Jung et al. (2020)

		Used in: Final and structural cost estimation models
12	Number of basement floors	**Parameter Type**: Numerical; **Authors**: Dursun and Stoy (2016); Wen-der and Skibniewski (2010); Gunaydın and Dogan (2004) **Used in**: Final and structural Cost estimation models
13	Average floor height	**Parameter Type**: Numerical; **Authors**: Dursun and Stoy (2016) **Used in**: Final Cost estimation models
14	Type of External Finishing	**Parameter Type**: Categorical; **Authors**: El-Sawalhi (2015); Sawalhi (2012) & Sawalhi (2012) **Used in**: Final Cost estimation models
15	Formwork Type	**Parameter Type**: Categorical; **Author**: Wang et al. (2017) **Used in**: Final Cost estimation models
16	Location	**Parameter Type**: Categorical; **Author**: Dursun and Stoy (2016) **Used in**: Final Cost estimation models
17	Ground floor area	**Parameter Type**: Numerical; **Author**: Pal et al. (2015) **Used in**: Final Cost estimation models
18	Generator	**Parameter type:** Boolean **Author:** Oldfield (2009) identified the cost of a backup generator can range from 0.5% to 2.5% of the total building cost. Although it is necessary for all buildings to have a backup generator in some countries, from the data it was noticed not all buildings have generators in the BOQ and as a result those buildings with a generator have a greater final cost.
19	Typical floor area	**Parameter Type**: Numerical; **Author**: El-Sawalhi (2015); Sawalhi (2012); Pal et al. (2015) **Used in**: Final Cost estimation models

20	Building Function (type)	**Parameter Type**: Categorical; **Authors**: Vasily and Edward (1974); Pasquire (1999) **Used in**: Final Cost estimation models
21	Number of households	**Parameter Type**: Numerical; **Authors**: Kim et al. (2004); Jung et al. (2020); Pal et al (2015) **Used in**: Final Cost estimation models
22	Ground Condition	**Parameter Type**: Categorical; **Authors**: Dursun & Stoy (2016) **Used in**: Final Cost estimation models
23	Soil Type	**Parameter Type**: Numerical; **Author**: Wang et al. (2017) ; **Used in**: Final Cost estimation models
24	Soil Bearing Capacity	**Parameter Type**: Numerical; **Author**: Idowu (2019) **Used in**: Structural work quantity estimation models
25	Total Building height	**Parameter Type**: Numerical; **Authors**: Idowu (2019); Yeh (1998) **Used in**: Structural work quantity estimation models
26	Gross building Volume	**Parameter Type**: Numerical; **Authors**: Dursun & Stoy (2016) **Used in**: Final Cost estimation models

2.5. SUMMARY AND GAP IDENTIFICATION

Although numerous studies have been conducted on the topic, these techniques have yet to be broadly accepted in the industry. The lack of standardization in the production of the models is one of the reasons for this. While the bulk of these experiments rely on the usefulness of a single machine learning or predictive technique, such as ANN, CBR, or Regression, in forecasting costs, they tend to take very different approaches to designing and testing their models. There does not seem to be a systematic method for creating and testing predictive models, including those that are identical. This lack of standardization can be found in majority of the researchers' methodologies. Aside from Oluwafunmibi & Lam (2020), the remaining refered works did not adopt any standardized methodolgy for data mining, whereas some tried

to standardized their own techniques. As noted by García de Soto (2014), while lack of standardization can be acceptable for research, this needs to change if these techniques are going to be used by practitioners in the industry. This gap can be addressed by adopting a renowned and well accepted Industry standard methodology for data mining.

It was also noticed that these researches employed qualitative methods such as questionnaire surveys and interviews to identify and rank the important predictor variables to use in their models. But these approaches are subjected to personal thoughts and does not quantify the relationship between the predictor values as well as betweeen the predictor value and the output. Adopting a quantitative feature selection technique to identify and rank the predictor values prior to using them in the prediction models have been tested to increase the accuracy of prediction in other researches - case in point the study by Kumar et al. (2013). But it is yet to be tested for the estimation of building project costs in Ethiopia.

On a related note, the quantity-based approach for cost estimation requires to first predict the quantity of works for different work items, multiply them with their respective unit rates to get the cost for the individual work items and then sum these costs to determine the final costs. Individually, these quantity of works predictions come with their own uncertainty. Summing these uncertainties will result in the case of error propagation. In the literature review, it was noted that none of the researches that compared the quantity and cost-based approaches accounted for this propagation of error in their comparison. This gap will be dealt in the following research.

Lastly, considering the price flucuations seen in Ethiopia, Literature suggests quantity based approach would provide a more accurate result (Joseph, 2013), but this has not yet been researched and tested for the case of Ethiopia. In fact, none of these data mining techniques have been tested and no research has been done regarding the potential use and performance of these techniques in the Ethiopian building construction industry.

3. RESEARCH METHODOLOGY

The methodology adopted for the proposed work is one called The Cross-industry standard platform for data mining (CRISP-DM). The CRISP-DM is a widely accepted and adopted methodology for data mining projects (Rudiger & Jochen, 2000). This methodology involves six stages and will be described as they are planned to be used in this research.

3.1. Stage 1 – Business Understanding

The first stage focuses on having a clear understanding on the project objectives and to have adequate knowledge of the subject matter so that the problems are clearly stated and justified. This stage has a qualitative approach to it. This entails doing an extensive literature review on the researcher side to gain as much background information and knowledge regarding the subject matter.

Accordingly, a significant amount of time and effort has been invested to identify and study different study materials such as books, journal papers, conference proceedings and graduate and doctoral theses that are related to the subject matter. Furthermore, it is at this stage the kind of data that is needed is identified. Which, for the case of this research, includes identification of factors that affect the final cost, structural cost and structural work quantities of building projects.

These factors need to be ones that are known at the preliminary stage of a project; where the design for the building is not prepared but some information is known regarding the general characteristics of the building. With that in mind, a total of 26 factors are first identified from literature review (discussed in section 2.4.1**)** but considering the availability of data (or lack of) for some of the variables such as: *duration, soil type, soil bearing capacity and Gross soil reaction*; similarity amongst some of the features *like total slab area Vs. gross floor area Vs. typical floor area vs. ground floor area* as well as location, internal wall finish and formwork type because the data was collected for buildings in Addis Ababa and had similar specification for formwork as well as similar finish material for internal walls, the number of predictors was filtered down to 16 predictors for final cost prediction model.

These 15 predictors were further filtered with regards to their (un)importance for structural work quantity or cost estimation (predictors related with finishing works as well as electrical works) and 9 predictors were chosen for the structural cost and quantity prediction models. Table 6 summarizes the different features and on which models they are to be used.

Table 6: Models and the Respective Attributes to Be Used for Prediction

No.	**Feature**	**Final cost**	**Structural cost and quantity estimation models**
1	Slab type (ST)	Y	Y
2	Foundation type (FT)	Y	Y
3	Total Slab area (TSA)	Y	Y
4	Basement floors (BST)	Y	Y
5	No. of floors (NF)	Y	Y
6	Slab thickness (SLT)	Y	Y
7	No. of lifts (NL)	Y	Y
8	Average floor height (AFH)	Y	Y
9	Shoring work (SW)	Y	Y
10	Building type (BT)	Y	X
11	Internal Decoration (floor) (IDF)	Y	X
12	External Decoration (ED)	Y	X
13	Generator (GNR)	Y	X
14	Fire System (FS)	Y	X
15	Data system (DS)	Y	X

Y ---- used, X ----- not used

3.2. STAGE 2 – DATA COLLECTION & INVESTIGATION

The second stage consists of data collection and investigation. Supervised predictive models are heavily reliant on the data on which they are trained on. Though the algorithms chosen plays an important role on the performance of the models to be built, the quality of the data used to train these models is just as important. Using poor, unverified data will result in a case of Garbage-in-garbage-out (GIGO), where the prediction we gain from these models will be poor and do not represent the actual situation.

Sampling Techniques

The sampling technique adopted for the data collection is Purposive sampling technique and the criteria for the sample is building projects, residential as well as commercial built in Addis Ababa within the past ten years that were built on a unit price contract. The data required to develop these predictive models is gathered from contract and design documents, particularly from payment certificates, as-built drawings, bill of quantities, and takeoff sheets. The data is collected from private and government consultants, contractors and clients.

Determination of Sample size

From a statistical perspective, in order to develop a model and validate its performance, it is necessary to have a certain minimum number of instances (samples) that can be a representative of the population. However, identifying the samples size can be difficult, especially when the population is large (Oluwafunmibi & Lam, 2020).

Most machine learning algorithms, especially the non-linear ones, benefit from large data as smaller data sets tend to be sensitive to noise and highly subjected to bias. Unfortunately, finding more data is not an easy task. In many instances, the data needed is not properly documented and saved and as a result, it would require a significant amount of investment, both in time and money, to increase the data set. This creates a sort of trade-off between the cost of underperformance as a result of having less data and the cost of collecting more data. However, even though there are many recommendations, equations and practical rule of thumbs, there seems to be no accord on a single approach to identify the minimum sample size.

In Table 7, a summary is presented outlining some of the famous rule of thumbs used to identify the minimum sample size. As it can be seen, the number of sample size is highly dependent on the number of feature or attributes to be used in the models.

Table 7: Rule of Thumbs to Identify Minimum Sample Size

Author/s	Recommendation
Nunnally (1978)	The minimum ratio of the sample size to the number of attributes should be 10
Kass and Tinsley (1979)	The minimum sample size should be 5-10 times the number of attributes
Comrey and Lee (1992)	Below 100 is poor, 300 is good and 1000 is excellent

Though these recommendations are based on experience, there seem to be no mathematical backing to any of them. Garcia De Soto (2014) proposed a mathematical equation that takes into account the number of independent variables, coefficient of determination and train-test split ratio to determine the minimum number of instances required to develop a model.

For this research, the equation proposed by García de Soto (2014) is adopted to estimate the minimum data required for these models. The least desired coefficient of determination, R^2_{min}, implies how much of the variation in the data we want to be described by the model, an R^2 value of 0.9 means 90% of the variation in the independent variables can be described by the model. But this would mean that R^2 value will always increase with an increase in features as more features will mean more information gain but in reality, some features may not help in explaining the variation in the model and even worse, it will increase the dimensionality of the data resulting in over fitting.

To remedy this problem, another metric known as Adjusted R^2 given by equation 11 is presented:

$$Adj\ R^2 = 1 \quad \left(\frac{n+k+1}{n-k-1}\right)(1 \quad R^2) \qquad (11)$$

Where: 'n' implies sample size and 'k' is the number of independent variables/features

Unlike R^2, this metric only increases when the independent variable added has a significant effect on the dependent variable (García de Soto, 2014). Using the ratio of the adjusted R^2 and R^2_{min} in the equation, a minimum number of data instances required can be calculated using the equation

$$L(n) = \frac{(k+1)(\ R^2_{min}\ \ Z+R^2_{min}-2)}{S\ \ R^2_{min}(Z-1)} \qquad (12)$$

Where:

- *K is maximum number of features to be used in the model (to be assumed).*
- *R^2_{min} is the least desired coefficient of determination (0.9),*
- *Z is the ratio of Adj.R^2 to R^2_{min} (recommended 0.9) and*
- *S is percentage of data to be trained and tested (as it is planned to use 10-fold Cross-validation, the train-test split ratio will be considered as 90% training and 10% testing repeated for each 10-fold. i.e., in each fold, 90% of the data will be used for training and the remaining 10% for testing. This split ratio will be used for all 10 folds.* (García de Soto, 2014)

Using this equation and the number of features identified earlier, the minimum data required for developing the models can now be estimated.

- Accordingly, for the final cost model, this amounts to:

$$L(n) = \frac{(15+1)(0.9\quad 0.9+0.9\quad 2)}{0.9\quad 0.9(0.9\quad 1)} = 57.2\ or\ 58\ data\ points.$$

- For the remaining structural work quantity and cost models, $k = 10$ and $L(n) = 39.3, say\ 40\ data\ points$.

In case where $L(n) < N(actual\ collected\ data)$, then measures will be taken to collect more data. If that is not possible, then steps will be taken to reduce the number of independent variables using correlations as well as feature engineering method like step wise selection, which uses forward selection and backward elimination along with Genetic algorithm to identify the most important features.

Once the data is collected, the methodology requires that the collected data to be studied. This means the data will be investigated to gain insight into any unique relationships within the data

sets and to study the quality of the data and find hidden information within the data and will be reported.

3.3. STAGE 3 – DATA PREPARATION

- This stage is called the data preparation stage and it is where the data will be prepared so that the model can be constructed.

  For this research case this entails:

  - Data organizing: sorting through the data and organizing the ones needed in an excel sheet.
  - Converting categorical data like slab type, foundation type, finishing works to numerical.
  - Adjusting costs for inflation: As noted in Stage 1, the data is collected for buildings built since 2010. As the value of money change with time, the cost data for these buildings will need to be deflated or inflated and be expressed in terms of a similar year value, thereby avoiding the effect of inflation. This is done by adjusting the costs to the year 2019. The Consumer Price Index (CPI) is used to adjust these costs for inflation. The "ideal" CPI for this case would be a construction CPI where the commodities used are construction specific. As no data on construction CPI can be found, a general consumer price index was used to adjust the costs for inflation. The CPI data was collected from Central statistics agency and is presented in Table 8.

    The equation used for the adjustment is presented as follows:

    $$Cost\ in\ year\ Z\ = cost\ in\ year\ X \quad (\frac{CPI_x}{CPI_z}) \qquad (13)$$

    Where, x and z are specific years (z being year 2019 for this case)

  - **Outlier Detection:** outlier search was carried out using the outlier detection approach recommended by Ramaswamy, Rastogi and Shim in "Efficient Algorithms for Mining Outliers from Large Data Sets". In their paper, a formulation for distance-based outliers is proposed that is based on the distance of a point from its ***k-th*** nearest neighbor. Each point is ranked on the basis of

its distance to its ***k-th*** nearest neighbor and the top ***n*** points in this ranking are declared to be outliers. After trying different combinations of k and n in a regression model, the value of ***k*** was taken as 1, and the value of n was set to be 6 as that dataset provided a more accurate regression model.

Table 8 Consumer Price index for years 2010-2019 (source: Worlds Bank and Central Statistics Agency)

Consumer Price index for years 2010-2019 taking 2010 as a base year.	
Year	CPI
2010	100
2011	132.0148
2012	162.878
2013	175.0352
2014	187.0952
2015	204.9981
2016	218.5857
2017	241.9344
2018	275.3958
2019	319.0194

- **Normalizing the data:** Some algorithms like that of Neural networks, particularly those using sigmoid activation function, require the data to be normalized so that each feature or independent variable with different scales are converted to a similar scale allowing all variables to get treated fairly (i.e., variables with large values will not become dominant). The normalization is done automatically in the RapidMiner® software using a range linear transformation.

3.4. STAGE 4 - MODELLING

In the modeling stage, modeling algorithms or techniques are chosen and applied, and their parameters are calibrated until the model's accuracy cannot be improved any further. For this research, 4 modeling techniques' performance is assessed for the different models developed. The reasons for selecting these techniques are presented as follows:

a. **Linear Regression:** In prediction modeling, it is always good to begin with the simplest, least complex model. And it doesn't get simpler than liner regression. Linear regression allows us to study if there is any linear relationship between the predictor variables and the prediction.

b. **Decision Tree:** DT is a supervised machine learning model that uses a repeated splitting algorithm to break cost data into hierarchical rules on each tree node (Breiman, et al., 1984). Decision Trees are among the most popular machine learning algorithms used in different areas. Their interpretablity has given them advantages in most classification problem. Decision trees have an added advantage of handling non linear data when compared with Linear regression.

   When Decision Trees are used for regression problems, they are called regression trees and though researches like Shin (2015) has determined their superiority with regards to interpretability as well as performance as compared to LR or NNs, there seems to be little research to follow back on his study. Furthermore, as Decision Trees are base learners for Gradient boosted trees, the performance of a single decision tree was needed to compare with a boosted version of it. Decision Trees will be built using the least square criterion which dictates splitting of nodes shall improve the fitting error of the resulting tree and the error for the node t is given by:

$$Err(t) = \frac{1}{n_t}\sum_{dt}(y_i \quad k_i)^2 \qquad (14)$$

$w\ \ ere\ y_i\ is\ t\ \ e\ actual\ value\ and\ k_i\ is\ t\ \ e\ predicted\ value\ at\ node\ t$

- In order to arrive at the most accurate decision tree model, different combination of the parameters for the decision tree model like the maximum number of trees, and other stopping criteria like the minimal gain or maximum error for a node to have in order to split and minimum size of the node for split

will have to be tried. As doing this manually takes too much time, a grid optimization technique was used. A Grid optimization uses the ranges of the parameter values and develop the model based on the different combinations of these parameter values.

c. **Neural Network:** Neural Networks are one of the most researched techniques for estimating preliminary costs and quantities. This might be owing to their consistent higher performance seen in several studies. As a result, Artificial Neural Networks were evaluated as another technique in this study.

The neural network architecture used in this thesis is a multilayer perceptron (MLP). A MLP is a class of feed forward neural network where connection between the inputs and hidden layers is directed only forward and the model is trained by a back-propagation algorithm. An MLP is made up of many layers of nodes in a directed network, with each layer completely linked to the one before it. Each node, with the exception of the input nodes, is a neuron (or processing element) with a nonlinear activation function, which in this case is the sigmoid activation function.

- The neural network's accuracy depends on the:
    i. Training cycle, also known as epoch (the number of times the model is trained, its prediction compared with the actual value and the error is fed back through the network). Too few values will result in poor generalizability and too many will result in too much computation time and over fitting.
    ii. Learning rate: the rate by how much the weights for the neurons should change. Too small will result in longer computation time and too large results in unusual divergence in the error calculation.
    iii. Momentum: aids the gradient descent algorithm in not getting stuck in local minima.
    iv. Number and size of hidden layers: the number and size of hidden layer describe the complexity to the problem at hand.

With regards to the number of hidden layers, literature suggests that one hidden layer is mostly enough for majority of the problems and thus are good starting points (Heaton, 2017).

Regarding the size of the hidden layer, there are many rule-of-thumbs that can be used for determining the correct number of neurons to use in the hidden layers, like:

- The number of hidden neurons should be between the size of the input layer and the size of the output layer.
- The number of hidden neurons should be 2/3 the size of the input layer, plus the size of the output layer.
- The number of hidden neurons should be less than twice the size of the input layer (Hornik, 1991).

These three rules only provide a starting point to consider and the selection of the number and size of hidden layers for the models, ultimately comes down to trial and error. Accordingly, the number and size of the hidden layers were found through trial-and-error basing on the rule of thumbs discussed above. Parameter optimization using Genetic Algorithm was carried out to identify the values for the remaining 3 parameters discussed above.

d. Gradient Boosted Trees: In the field of data science, gradient boosted trees and its derivations have lately been the go-to approaches. Their ability to compare and even surpass far more sophisticated deep learning models, as well as the simplicity with which improved models can be constructed with minimal hyper parameter adjustment, has made them popular in recent years (Thomas, Coors, & Bernd, 2018).

The application of these techniques in conceptual cost estimation has recently begun to be studied, and as discussed in the literature review, the GBT models were seen outperforming ANN as well as other models. To this end, GBT are chosen as the last technique to be evaluated for this research.

The base learner adopted for the gradient boosted tree models are decision trees. The process for building the models follows a similar approach to the rest of the models with the other three algorithms.

- The GBT model has a number of parameters that need to be adjusted to produce an accurate model. Trying to tune them with a brute force grid optimization technique takes too much time as it would be trying too many combinations. Thus, first, recommended values and ranges of values were identified through literature review and based on that, a hyper parameter optimization was carried

out using Genetic algorithm. The values gained from the parameter optimization were then tested using different runs to make sure the optimization was not stuck at a local optimum value.

The recommendations adopted for each parameter from literature review are presented as follows.

- Learning rate and number of trees: the smaller the learning rate, the larger number of trees required for similar performance. Friedman (1999), the father of Gradient boosted trees, in his paper Stochastic Gradient Boosting, recommended to first set a large value for number of tress (range between 100 and 500) and then tune the learning rate starting with 0.1 and lower to achieve best results.
- Sampling Percentage: according to Friedman (1999), the sampling percentage can take a range of 0.3 to 0.8 depending on the problem.
- Number of terminal nodes: with regards to number of terminals, smaller values like 3 or 6 are better than large values (Jason Brown, 2020).
- Number of Nodes in tree: based on some empirical studies, Trevor Hastie (2017) recommends the number of nodes (J) be between 2 and 10, with number of tree value of $4 <= J <= 8$ working well in most cases.

A total of 4 different predictions were made using the four data mining techniques discussed above – making the total number of models developed 24. The four predictions that are done are the following:

- **Final Cost Prediction models**: To predict the final cost based on historical final cost data of the buildings
- **Concrete Quantity Prediction Models:** To predict the quantity for concrete works including sub and superstructure concrete works.
- **Rebar Quantity Prediction Models:** To predict the quantity for reinforcement works including sub and superstructure reinforcement works.
- **Formwork Quantity Prediction Models:** To predict the quantity for Formwork (shuttering) works for sub and superstructure.

- **Structural Cost Prediction Models:** To Predict the Cost for Structural works (including concrete rebar and formwork) using historical cost data of buildings. This model is required to assess the effect of propagation of error that occurs as a result of using quantity models.

3.5. STAGE 5 - EVALUATION

- The **fifth stage** is Evaluation. It entails assessing the performance of the built or trained models by testing them on an unseen data set. This process is also called model Validation. The purpose of validation is two folds:
  - Primarily, Validation is used to evaluate the generalizability of the trained model by assessing how the model perform on an independent data set. This is part of the model building stage as validation is used to tune in parameters of the models.
  - When comparing different models, Validation is also used to identify or rank the models based on their performance.
- A model can be validated in one of two ways. The first is referred to as holdout validation. It entails withholding a portion of the data gathered so that it can only be used for assessment (testing). A 70-30 or 80-20 train-test split is commonly used to divide how much of the data can be used for training and testing. The problem with this approach is that it necessitates a larger data set as 20-30% of it would be disregarded during model training. Furthermore, since it divides the data sets at random, the output result obtained could be solely by chance, and the result could differ substantially if the data were split differently.
- The second choice is to use K-fold cross validation, which divides the data set into K folds or partitions and trains the model k times, using the first k-1 partitions as a train data set and the remaining one-fold as a test partition. The output of each iteration is summed to get the model's performance, which is presented as mean +/- standard deviation after repeating this process K times. This allows each data row to be used once for training and once for testing, which is useful when there is a lack of data. It also helps to avoid over fitting.
  - Considering how data thirsty holdout validation is and the small data set we have, a 10-fold cross validation is chosen for this research.
  - In order to attain producible result, a local seed was used to make sure that the data is partitioned exactly the same way for every model.

- Performance Metrics

  There are numerous performance metrics that can be used to evaluate a prediction. Table 9 identifies some of the most adopted metrics from literature review.

Table 9: Performance metrics adopted in different researches

Author/Year	**Models Used**	**Evaluation techniques used**
Elmousalami (2019)	FL, ANN, MRA, CBR, Hybrid techniques	MAPE, R square
Tadesse and Dinku (2017)	ANN	MSE (Mean squared error)
N.Bhirud and Vinayak (2017)	ANN	Mean Percentage error
Wang et al. (2017)	FALCON with GA	R square
Shehatto (2013)	ANN	Mean absolute error, MAPE, total MAPE, R square
Bayram et al. (2013)	ANN and RBF	Root mean Squared error (RMSE), MAE, R square
Oluwafunmibi and Lam (2020)	SVR	MAPE

- Based on literature reviews, the most common and important metrics in such predictions are identified. These are:
  - Average Absolute Error: given by $AE = \left(\frac{\sum_n |actual_n - predicted_n|}{N}\right)$
  - Root mean squared error: given by $RMSE = \sqrt{\frac{\sum_{i=1}^{N}(predicted - actual)^2}{N}}$
  - Average Relative Error or mean absolute percentage error: given by

$$\text{RE} = \left(\frac{100\ \left(\frac{\sum_n |actual_n - predicted_n|}{actual_n}\right)}{\text{N}}\right)$$

Where: N is the total number of rows to be tested and

n is the n^{th} row of the test set (Trevor Hastie, 2017).

Though the models are tested with all three metrics, the Relative error and mean absolute error are used to compare between models as the RMSE is sensitive to outliers and according to Wllmott & Matsuura (2005), it shall not be used as an error measure when comparing different algorithms.

SOFTWARE CHOICE

The RAPDIMINER® Software is used for data prepare developing these models as well as in data preparation. RapidMiner is an open-source data science software platform that is extensively used in data mining and machine learning applications. It is widely used in a number of business and commercial applications as well as in various other fields such as research, training, education, rapid prototyping, and application development. All major machine learning processes such as data preparation, model validation, results visualization, and optimization can be carried out by using RapidMiner. While it is an enterprise scale software, it offers a free scale for those in Academia.

Its friendly visual interface, ability to allow data exploration with powerful visualizations and develop models in a process flow chart system that can be easily understood by others with little knowledge of the software as well as its large community and support base has made it the first choice for this research.

3.6. STAGE 6 - DEPLOYMENT

- The **sixth and last Stage** is the deployment stage. Once the models have been built and validated, the knowledge gathered must be structured and presented. Furthermore, the models shall be deployed and be used to predict costs. (Rudiger & Jochen, 2000)
    - For the case of this research this means two things:
        - The First one is writing a report on the knowledge gained throughout the process and identifying which approach (the quantity based or cost based) along with the performance of each modeling techniques.
        - The second part is to deploy the best models so that others can use it for prediction. The models are deployed by creating an Application Programming Interface (API) that allows to serve the models into a web app that can display a clean User interface where users can interact with by imputing values for the independent variables and can see the output for the predicted variable.
        - This was done by exporting the models from the modeling software and importing to a server less cloud host called Clouderizer® to automatically create the API and build a web app with an input and output fields in the User interface. The web app can then be accessed through a web browser for scoring.

A summary of the research methodology is presented in Figure 4.

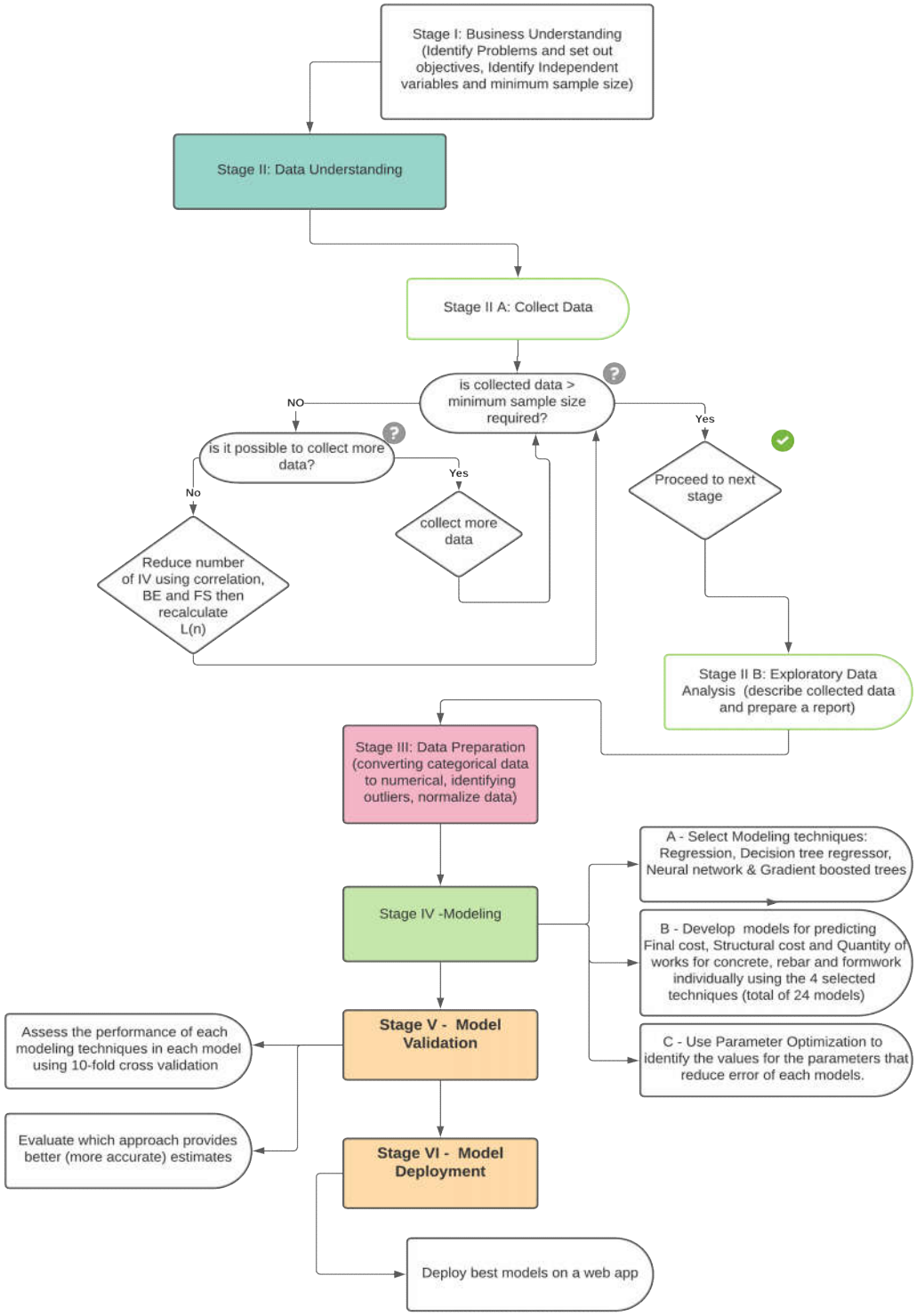

Figure 4: Research workflow

4. DATA DESCRIPTION AND PREPARATION

4.1. DESCRIPTION OF COLLECTED DATA

According to the Addis Ababa city administration infrastructure integration, construction permit and control authority, for the past 5 years, the city has annually issued an average of 35,000 – 40,000 building construction permits (IDCCPCA, 2020). Over the five year's period, this population amounts to two hundred thousand permits. But this is not the target population as this number includes buildings in all categories from small houses to high rise buildings. Furthermore, as some buildings are built on a labor – only contract, the final costs for such buildings are not well documented. Accordingly, the target population for this research are buildings having at least two floors including ground level that were built after the year 2010 and are constructed under a unit-price contract. This will significantly reduce the number of the target population but also makes it harder to estimate the exact number of population.

The sampling method adopted for this research is purposive volunteer sampling where the criteria for the data to be collected are as outlined in the target population. While public offices have the duty to disseminate data when requested, Private institutions are not obliged and thus, data from private institutions was collected based on their volunteerism.

The research is conducted by collecting 45 buildings' data from different public and private institutions. The data collected include design documents, Take off sheets, bill of quantities and payment certificates. Figure 5 shows the ownership distribution of the buildings.

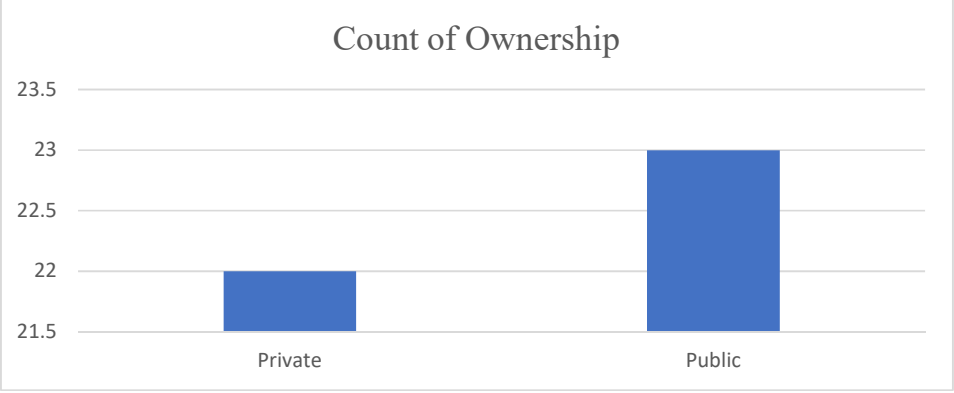

Figure 5: Ownership Distribution in the collected data

As it can be seen, the buildings have a somewhat equal distribution with regards to ownership, with 23 of the buildings having a public institution as an owner while the remaining 22 are owned by the private sector. The general description of the data with respect to each variable is presented as follows:

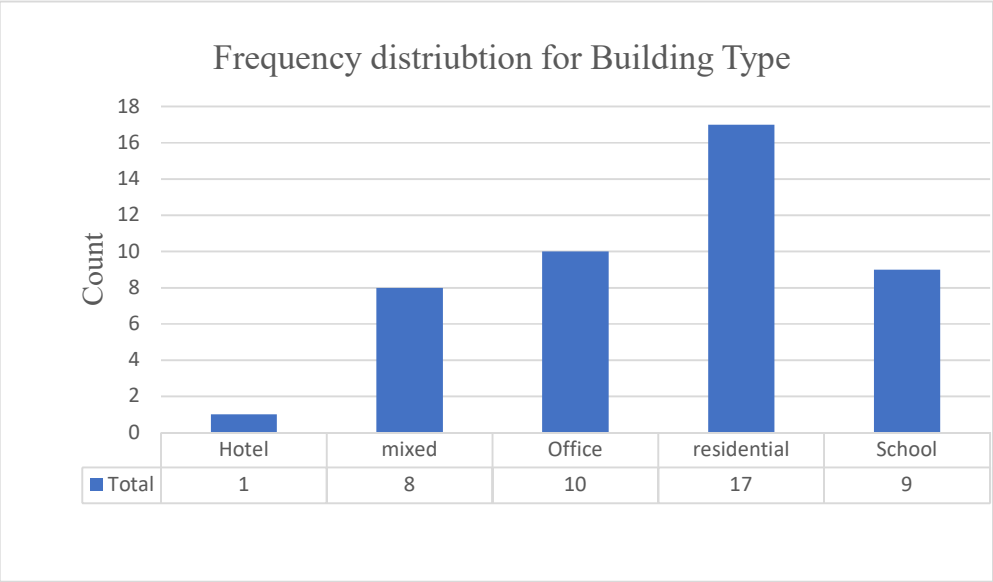

Figure 6: Frequency of Building Types in the Collected Data

Building type (BT)

The data consists of different building types including Residential, Mixed use, School, office and hotel buildings. The distribution of the building types is presented in Figure 6.

Type of Slab (ST)

The type of slab used in a building would influence both the cost of the building as well as the amount of concrete and reinforcement to be consumed. The building data collected is composed of two key types of slabs seen in the country: solid and ribbed slabs. The data consists of 18 buildings with ribbed slabs and the remaining 27 buildings have solid slabs.

Number of Basement (NB)

While majority of the buildings in the data do not have any basement, 10 of the buildings have a single floor basement and 3 have 2 basements underneath them.

Foundation type (FT)

The type of foundation is dependent on the type (and as the result, the load) of the building, type of soil, and neighboring structures. As these factors are mostly known at the conceptual stage, one can make the choice of the type of foundation to use comfortably. The data is composed of two types of foundation, namely, isolated footing and Mat Foundation. Majority of the buildings (36) are standing on an isolated footing while 9 have a mat foundation underneath.

Number of Lifts (NL):

28 of the buildings do not possess any elevators (even though some have floor stories way above the maximum story allowed to be built without any elevator). 9 buildings do have one elevator and 8 buildings have 2 installed.

Average Floor height (AFH)

The average floor height of the building is another variable used. The Average floor height is a variable gained by dividing the total building height with the number of floors. The buildings have average floor heights ranging from 2.38 meters to 4.66 meters. The frequency distribution for it is provided in Figure 7. As it can be seen, 62% of the buildings have an average floor height between 2.76 and 3.52.

Internal Floor Finishing (FF)

Different kinds of floor finishing materials are used in the buildings in discussion and have been categorized as follows:

- If more than 70% of the building floor are covered by PVC and/or terrazzo, then the building floor finish is grouped as Basic
- If more than 50% of the building floor are covered by Ceramic and/or the remaining floor is covered by PVC, the building floor finish is grouped as Standard.

- If more than 50% of the building floor are covered by Ceramic and the remaining floor is covered by porcelain, then the building floor finish is grouped as Luxury

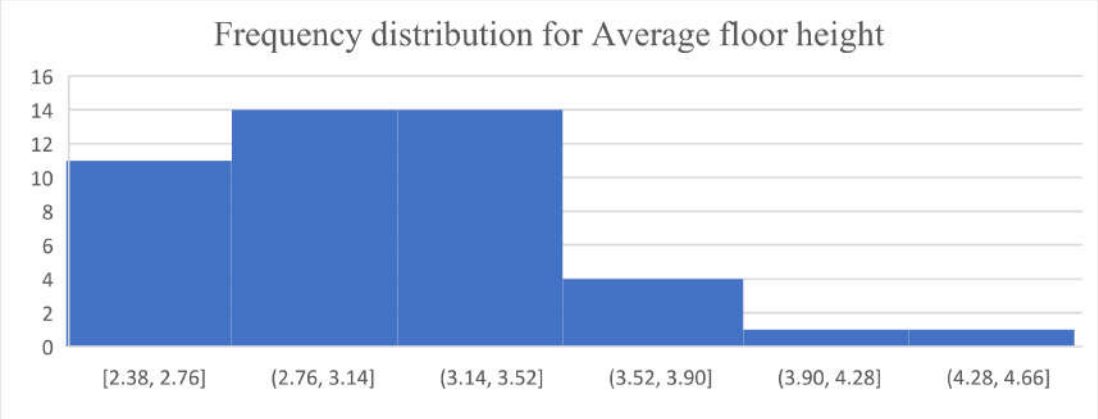

Figure 7: Frequency distribution of Buildings' Average height

Accordingly, 41 percent of the buildings have a luxurious floor finish while another 41% have a standard floor finish. The remaining buildings have a basic floor finish.

Total Slab area (TSA)

The total slab area is the gross building superstructure slab area. This includes the slab area for each suspended slab as well as the ground floor. The total slab area does not include basement slab area. The frequency distribution is provided in Figure 8.

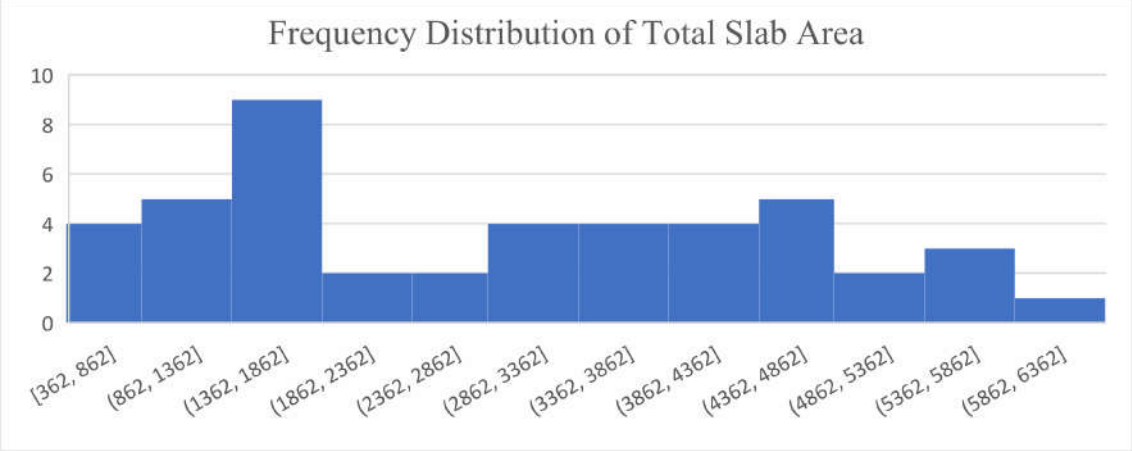

Figure 8: Frequency distribution of total slab area of the buildings

External Finishing Quality (EF)

External finishing quality entails any and all kinds of finishing done for the outer structure of the building. The finishing types used range from quartz paint on plastered wall to aluminum cladding walls. Based on the data collected, the finishing types are categorized as follows:

- Basic: synthetic paint applied on a plastered HCB wall
- Standard: quartz or Granite paint applied on a plastered HCB wall
- Luxury: Aluminum composite panel (ACP) cladding or semi-structural curtain wall. The frequency distribution for EF is presented in Figure 9.

Figure 9: Frequency distribution of External finishing quality in the data

Number of Floors

The number of floors indicated here implies the number of floors above the ground floor. As it can be seen from the frequency distribution (Figure 10), the number of floors range from one to eleven, with buildings with 3-4 floor heights having the largest count.

Shoring work (SW)

Some buildings may need permanent shoring work to protect them from lateral displacement of land. Shoring work becomes even more necessary when deep excavation is done for basements and foundations. The kind of shoring work in discussion here is a permanent retaining wall made of concrete. Out of the 45 buildings in question, 11 buildings have shoring works and all of them have at least one basement floors. This is understandable as basement requires deep excavation and according to Hailemariam et al. (2020) excavation depth, along

with location and geotechnical investigation are the major factors affecting the requirement for shoring construction in Addis Ababa, Ethiopia.

Slab Thickness

The thickness of the slab depends on the slab type. Ribbed slabs have a thickness of 28-30 centimeters while solid slabs have a thickness between 15-23 centimeters, of which, 78% are either 15 or 16 centimeters.

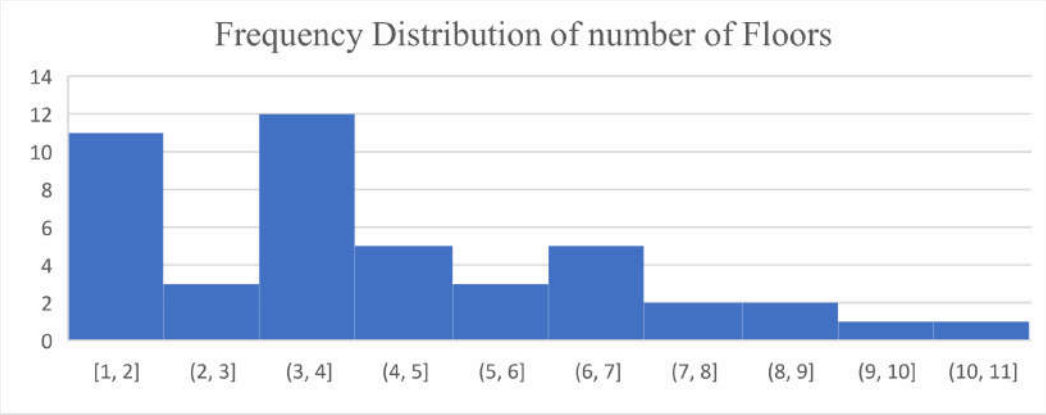

Figure 10: Frequency distribution of Number of floors in collected building data

4.2. EVALUATION OF SAMPLE SIZE AND FEATURE SELECTION

The data collected has 45 instances or data points. Based on equation 12, this sample was evaluated if it meets the minimum sample requirement size. While 45 data points is above the minimum requirement for the structural work cost and quantity and cost modes, it is still short of data points to meet the minimum requirement for a final cost model. In order to meet the minimum sample requirement, efforts were made to collect more data to but as it was not possible to collect more data in time, steps were taken to resort to another option. This option was to reduce the number of independent variables as much as possible based on the available data – a process called dimensionality reduction.

This was done by using correlation and stepwise feature selection - using regression as a base model. Before that, Average floor height and number of floors were first merged using multiplication to create a new attribute called average **building height**. The correlation for the discussed attributes is provided in Figure 11. As it can be seen from the correlation matrix, Slab type and slab thickness have a high negative correlation.

Attributes	Basem...	Shoring...	Building...	Building...	Data Sy...	Externa...	Fire Sys...	Floor De...	Foundat...	Generat...	Number...	Slab thi...	Slab Type	Total Sl...	Final Cost
Basement	1	0.840	0.631	0.387	0.045	0.321	0.375	0.299	0.467	0.348	0.594	0.208	-0.118	0.447	0.457
Shoring Work	0.840	1	0.641	0.322	0.145	0.225	0.323	0.241	0.595	0.234	0.586	0.047	0.127	0.497	0.476
Building Height	0.631	0.641	1	0.387	0.320	0.419	0.404	0.320	0.697	0.554	0.771	0.095	0.094	0.690	0.741
Building Type	0.387	0.322	0.387	1	-0.168	0.145	0.238	0.194	0.140	0.200	0.403	0.300	-0.191	0.206	0.190
Data System	0.045	0.145	0.320	-0.168	1	0.430	0.502	0.463	0.294	0.244	0.363	-0.004	0.178	0.418	0.419
External Finish	0.321	0.225	0.419	0.145	0.430	1	0.594	0.528	0.232	0.405	0.455	-0.011	0.225	0.373	0.417
Fire System	0.375	0.323	0.404	0.238	0.502	0.594	1	0.589	0.177	0.570	0.491	0.209	-0.026	0.298	0.320
Floor Decoration	0.299	0.241	0.320	0.194	0.463	0.5016881674068235		1	0.345	0.423	0.359	-0.035	0.191	0.283	0.313
Foundation Type	0.467	0.595	0.697	0.1		0.232	0.177	0.345	1	0.386	0.592	-0.044	0.217	0.452	0.492
Generator	0.348	0.234	0.554	0.200	0.244	0.405	0.570	0.423	0.386	1	0.545	0.038	0.071	0.299	0.307
Number of lifts	0.594	0.586	0.771	0.403	0.363	0.455	0.491	0.359	0.592	0.545	1	0.138	0.051	0.723	0.714
Slab thickness	0.208	0.047	0.095	0.300	-0.004	-0.011	0.209	-0.035	-0.044	0.038	0.138	1	-0.902	-0.019	-0.038
Slab Type	-0.118	0.127	0.094	-0.191	0.178	0.225	-0.026	0.191	0.217	0.071	0.051	-0.902	1	0.179	0.167
Total Slab Area	0.447	0.497	0.690	0.206	0.418	0.373	0.298	0.283	0.452	0.299	0.723	-0.019	0.179	1	0.975
Final Cost	0.457	0.476	0.741	0.190	0.419	0.417	0.320	0.313	0.492	0.307	0.714	-0.038	0.167	0.975	1

Figure 11: Correlation matrix for chosen attributes

It can also be seen that shoring work and basement floors have a rather high correlation (0.84). Multi-collinearity reduces the model's performance and fill it with redundant variables. For this reason, it is recommended to remove variables with high correlation. Senaviratna & Cooray, (2019) in their research regarding the effect of multi-collinearity in Regression models discuss that variables with a correlation value higher than 0.8 or 0.9 indicate a serious multi-collinearity problem but also stress that when dealing with small sample sizes, it is better to begin with removing those variables with the highest correlation and proceed with the modeling. This is important because correlation values only identify linear relationships and when it is planned to use non-linear models, removing variables on the basis of correlation alone will be costly as the non-linear relationship these variables could have to the dependent variable is not entertained. Thus, for this case, only variables with a correlation of 0.9 and above were dealt with by removing one of the variables, which for this case is Slab thickness and slab type.

To identify which variable to remove and which one to keep, stepwise feature selection was carried out using backward elimination and forward selection. In Figure 12, a bar graph is presented displaying the importance of the attributes. Those attributes recommended to be used are shown to have a weight of 1 while those not recommended have empty bars. Accordingly, variables Building Type, Fire System, Internal Floor Decoration, Number of lifts, Foundation Type, Generator and Slab Type are removed from the parameter pool.

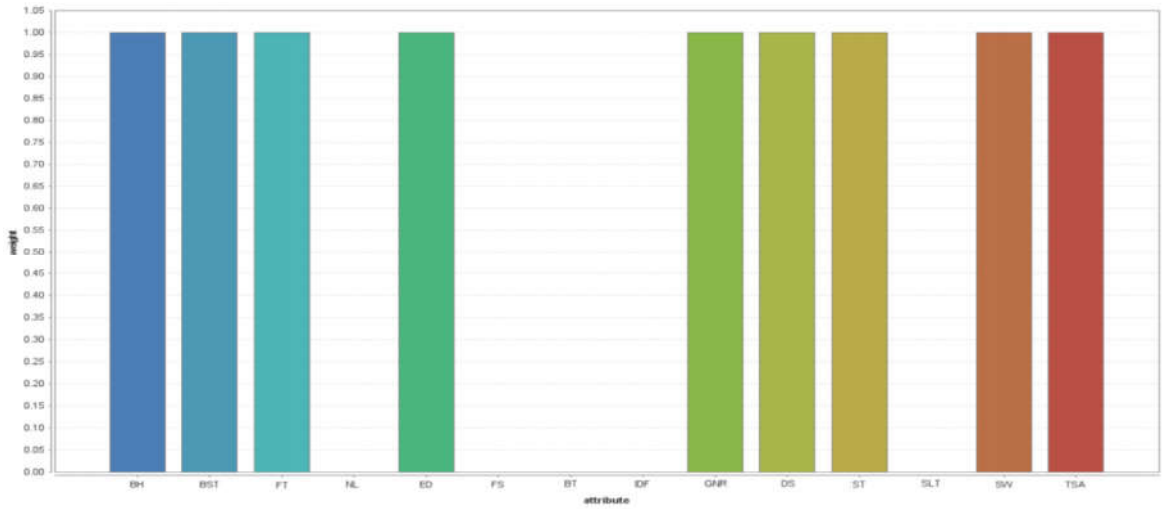

Figure 12: Attribute importance graph for final cost estimation (light blue – Least importance, Dark red – highest importance)

With the number of independent variables being equal to 9, this reduces the minimum sample size down to $L(n) = \frac{(9+1)(0.9\ \ 0.9+0.9-2)}{0.9\ \ 0.9(0.9-1)} = 35.8\ or\ 36\ points$ as the data at hand is more than the minimum, we can proceed to the next step.

4.3. DATA PREPARATION

Converting categorical variables to numerical

Categorical variables are converted using a label encoder. When the independent variable contains only two categories like that of slab type (containing only solid and ribbed slab), a value of 1 can be taken for one and 0 for the other. When the independent variable contains hierarchical values, like that of the quality of Internal finishing (containing basic, standard and luxury), a value of 1,2 and 3 are used signifying the level of luxury seen in the building internal finishing. The original and converted values are presented in Table 10.

The costs collected are also inflated to 2019 using equation 13. A sample calculation for one building is provided below.

$$Building\ ID: AFD14G4\ ;\ Year: 2014\ ;\ Cost\ in\ 2014\ =\ ETB\quad 54{,}085{,}671.88$$

$$\text{Cost of buildng in year 2019}\ =\text{cost in year 2014}\quad \left(\frac{\text{CPI}_{2019}}{\text{CPI}_{2014}}\right)$$

$$=54{,}085{,}671\quad \left(\frac{319}{187.1}\right)=\mathbf{92,2\ 14479\ }\boldsymbol{ETB}$$

When It is required to do so, as in the case of building a neural network model, the data is normalized between [-1,1] using linear transformation. The data set left after outliers were removed was still more than the minimum sample size and thus, can proceed with the modeling phases.

Table 10: Categorical variable values and their corresponding converted numerical values

Variable	**Categorical value**	**Numerical value**	**Variable**	**Categorical value**	**Numerical value**
Slab type	Ribbed	0	Internal finishing (floor)	Basic	1
	Solid	1		Standard	2
Foundation type	Isolated	0		Luxury	3
	Mat	1	External Finishing	Basic	1
Building type	Office	1		Standard	2
	School	2		Luxury	3
	Residential	3			
	Mixed Use	4			
	Hotel	5			

5. RESULTS AND DISCUSSION

5.1. DEVELOPMENT OF FINAL COST PREDICTION MODELS

One of the objectives laid out for this research was to develop Final Cost prediction model based on historical cost data of buildings. After the data is collected and necessary data preparation are carried out, different modeling techniques were used to come up with a best fit model. Among the Four modeling techniques chosen, the first one is the regression model.

5.1.1. REGRESSION MODEL

The process for developing the regression model is shown in Figure 14. After the data preparation process were carried out, 10-fold Cross validation was used to assess the performance of the regression model.

The best fit regression model had a rather large error with a relative error (MAPE) of 60.81% +/- 24.26% and an absolute error of 22,852,927.017 +/- 11,831,730.271. Both the bias and variance error are too high indicating there is no good linear relationship between the independent variables and the dependent variable. A scatter plot for the regression model is provided in Figure 13. As it can be seen, the model had also predicted some negative values, which in the case of estimating costs, does not make sense. This could mean that the data was too small for the regression model to extrapolate from.

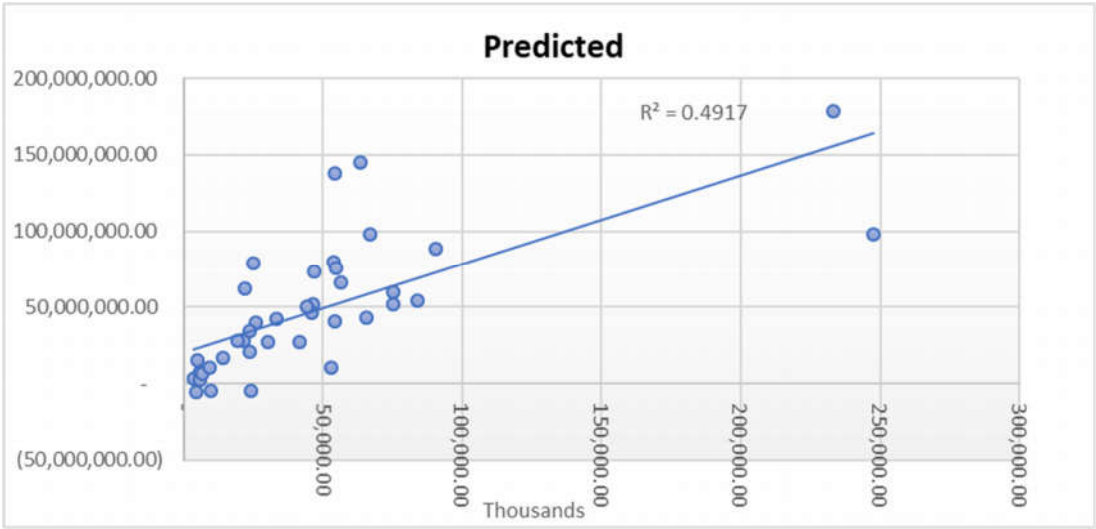

Figure 13: Best fit line for Final cost estimation regression model

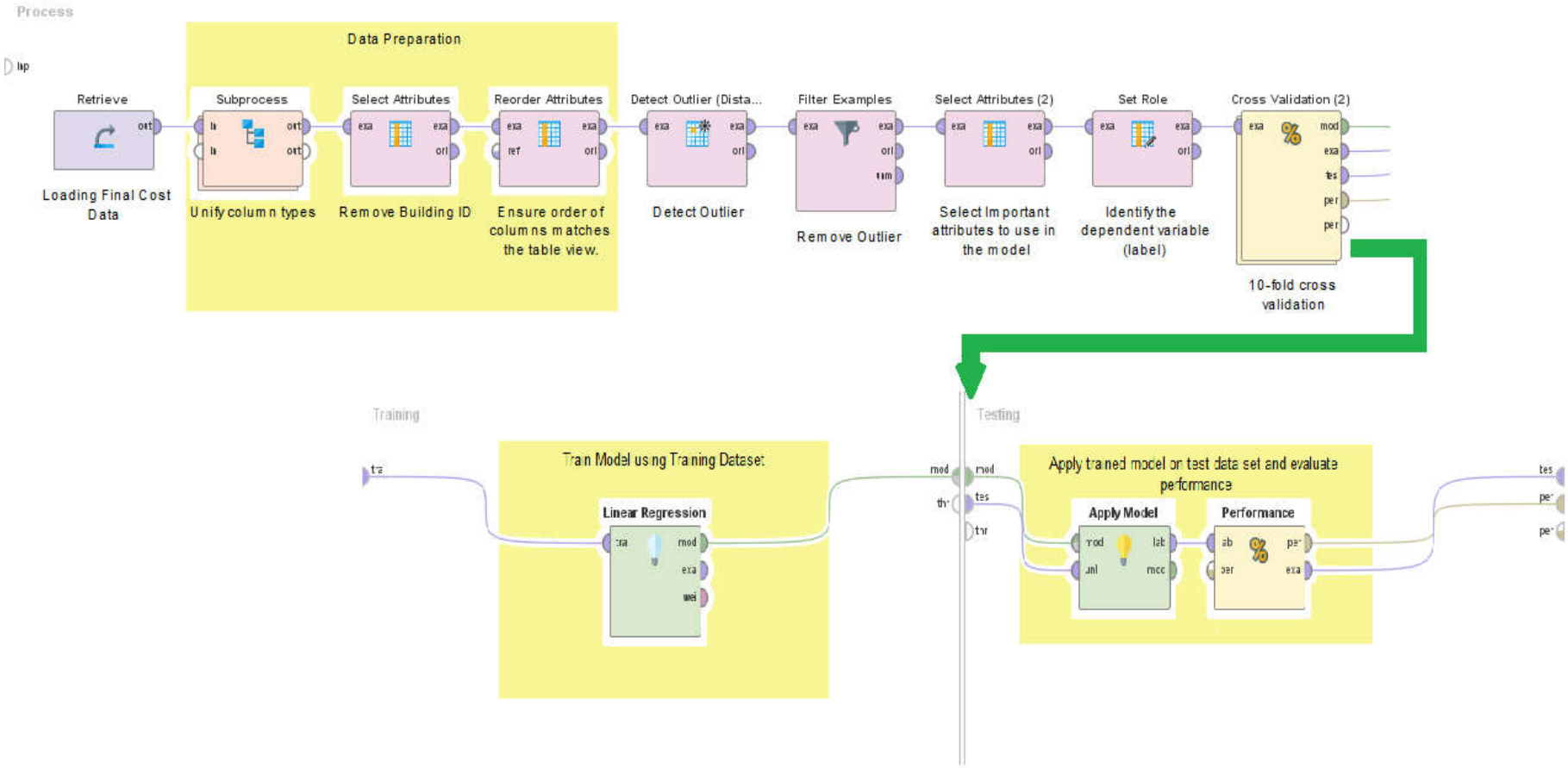

Figure 14: Process for developing regression Model for final cost estimation

5.1.2. DECISION TREE

The second algorithm evaluated is a decision tree. Least square method was used as a criterion for splitting a tree. In order to stop the tree from splitting for every instance and to reduce over fitting, a stopping mechanism was adopted in the form of preprinting.

The best decision tree model seems to outperform Linear Regression as both the bias and variance error has dropped considerably. The model has a relative error (MAPE) of 41.69% +/- 23.78% and an Absolute error of 17,524,247.337 +/- 16,265,246.926. The resulting decision tree as form of rule is provided in Figure 15.

5.1.3. NEURAL NETWORK

The neural network model process is shown in Figure 16. The Actual training and evaluation of the model is done inside the Cross-validation process which is shown underneath the cross-validation process block. The data preparation phase is similar to the remaining techniques with an exception of normalizing the data. Hyper parameter optimization was carried out to identify the ideal parameter values for the learning rate, training cycle and momentum whereas the optimum hidden layer number and size was determined through trial and error. A total of 12 different hidden layer number and size were evaluated. For each hidden layer number and size tried, a total of 22,220 combination of the 3 parameters were tried through automated parameter optimization taking 8+ hours on a dual core i7 2.4 GHz laptop. The best neural network model has 2 hidden layers with the first hidden layer having 15 nodes and the second one having 10 nodes.

The learning rate, training cycle (epoch) and momentum for the best model were 0.72, 382 and 0.3 respectively. In Figure 17, the learning rate vs. relative error graph is presented. As it can be seen, the relative error reduces with an increase in leaning rate until it reaches the global minima (at around 0.72 learning rate) then goes back up again. Indicating the global minima is at that point. It can also be seen that there are other places (around Learning rate of 0.2 – 0.3) where the relative error reduces a little bit only creating saddle points. These points are local minima values. The training cycle vs relative error shown in Figure 18 also shows similar results where the learning rate decreases up to some point then goes back up.

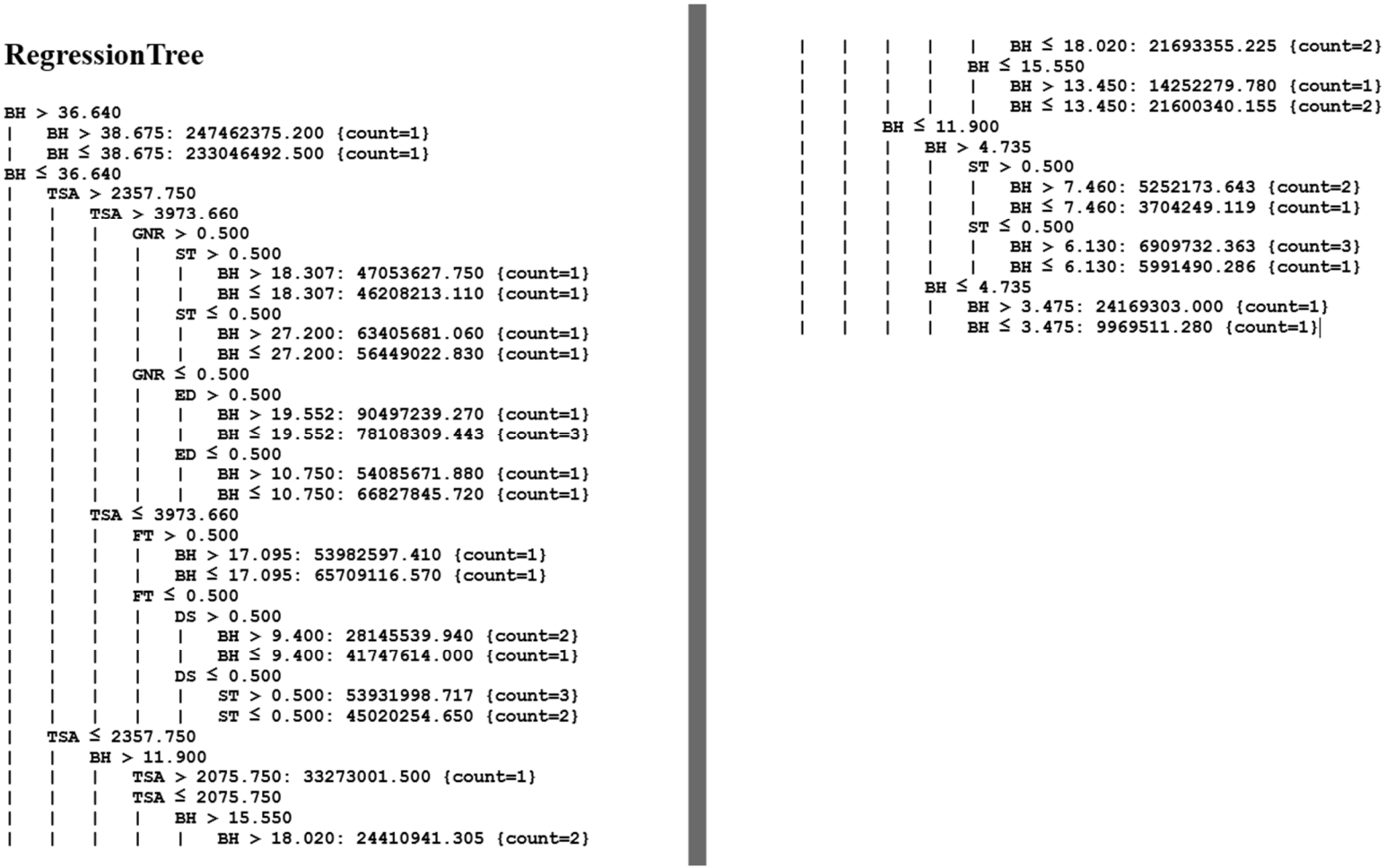

Figure 15: Decision Tree as a rule for Final Cost Estimation

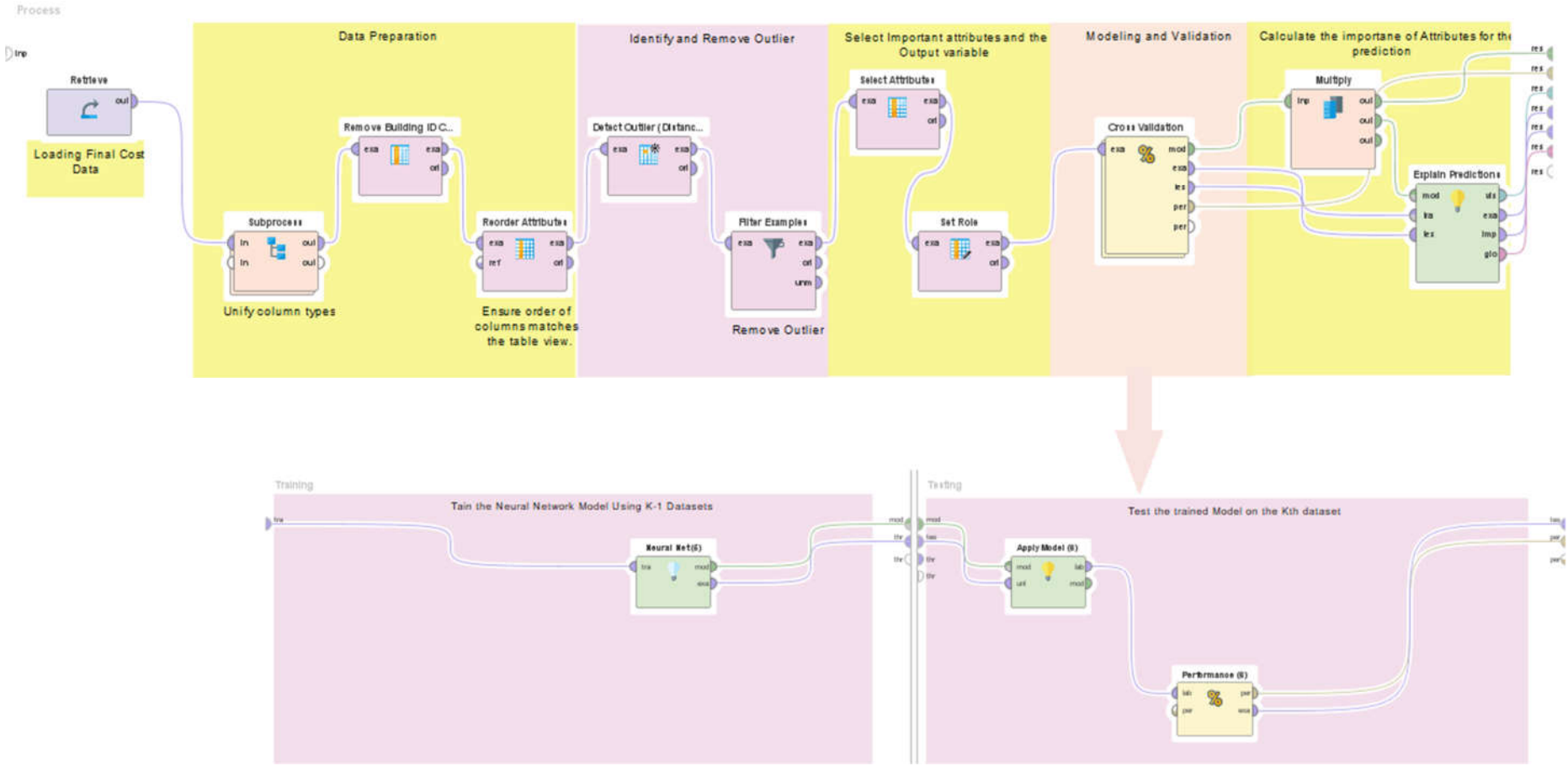

Figure 16: RapidMiner Process Model for final cost prediction using ANN

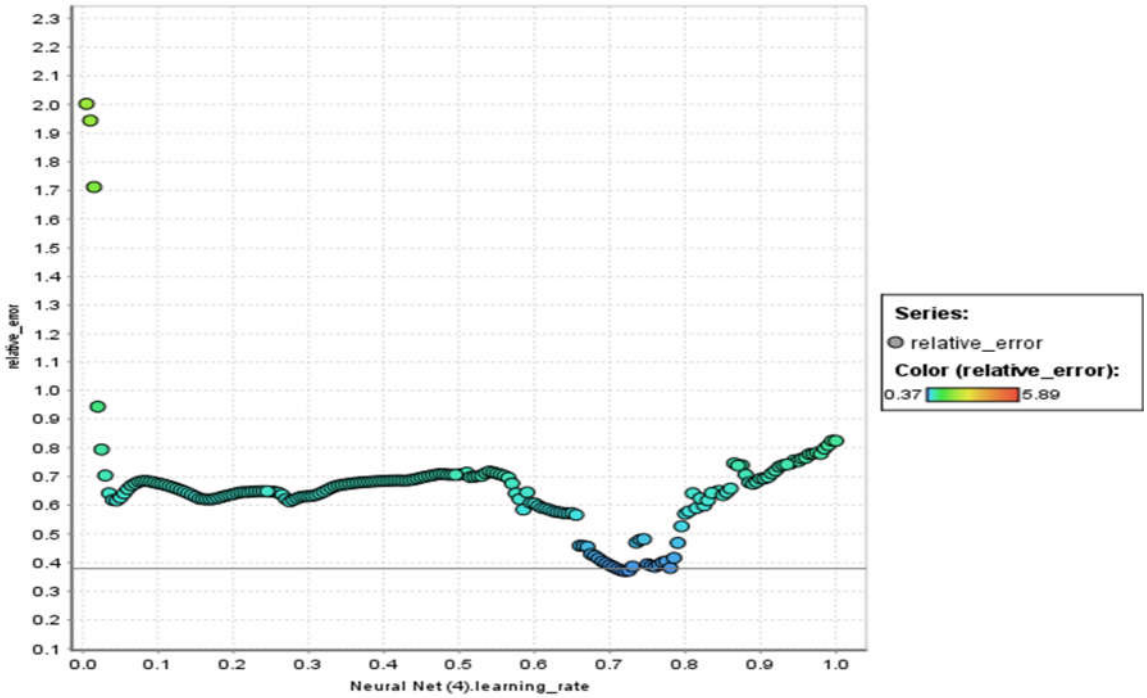

Figure 17: Learning rate Vs. Relative error during parameter optimization of NN model for Final cost estimation

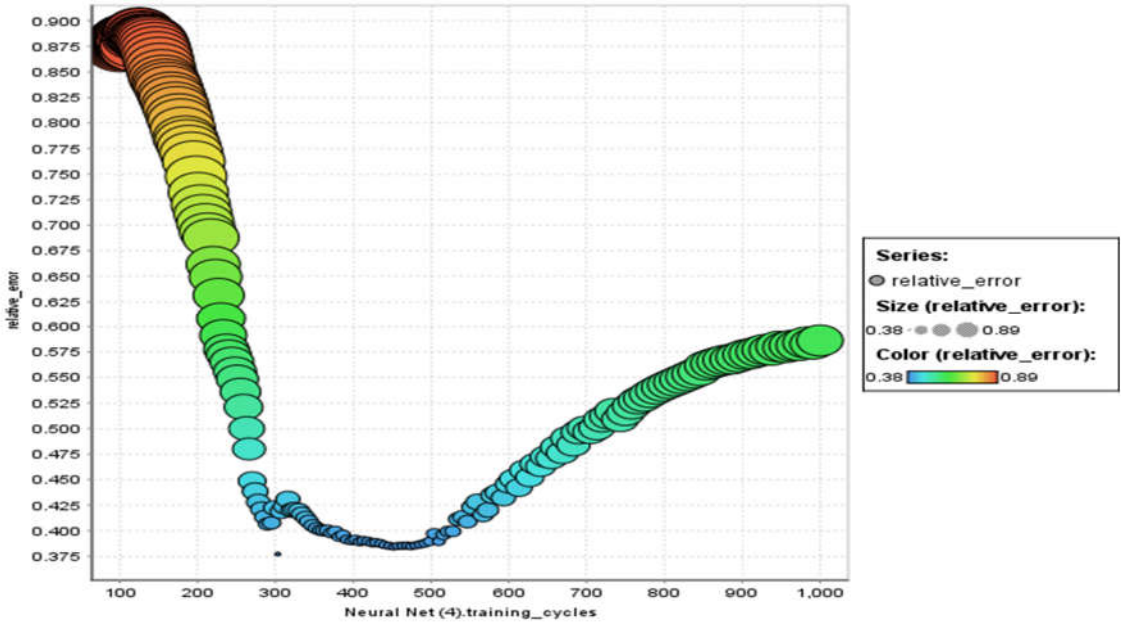

Figure 18: Relative error Vs. Training Cycles plot for Final cost NN Model

The neural architecture is presented in Figure 19. The performance of the Neural Network model was found to be slightly better than decision tree with the 10-fold cross validation producing an Absolute error of 14046002.432 +/- 6151162.912 and a relative error of 37.05% +/- 9.09%.

As it can be seen, though the bias error has only reduced by around 2 percent, the variance has also dropped considerably, indicating the model is more stable across different training sets. The weights for the input attributes as well as the nodes in the hidden layers are presented in Table 11 and Table 12.

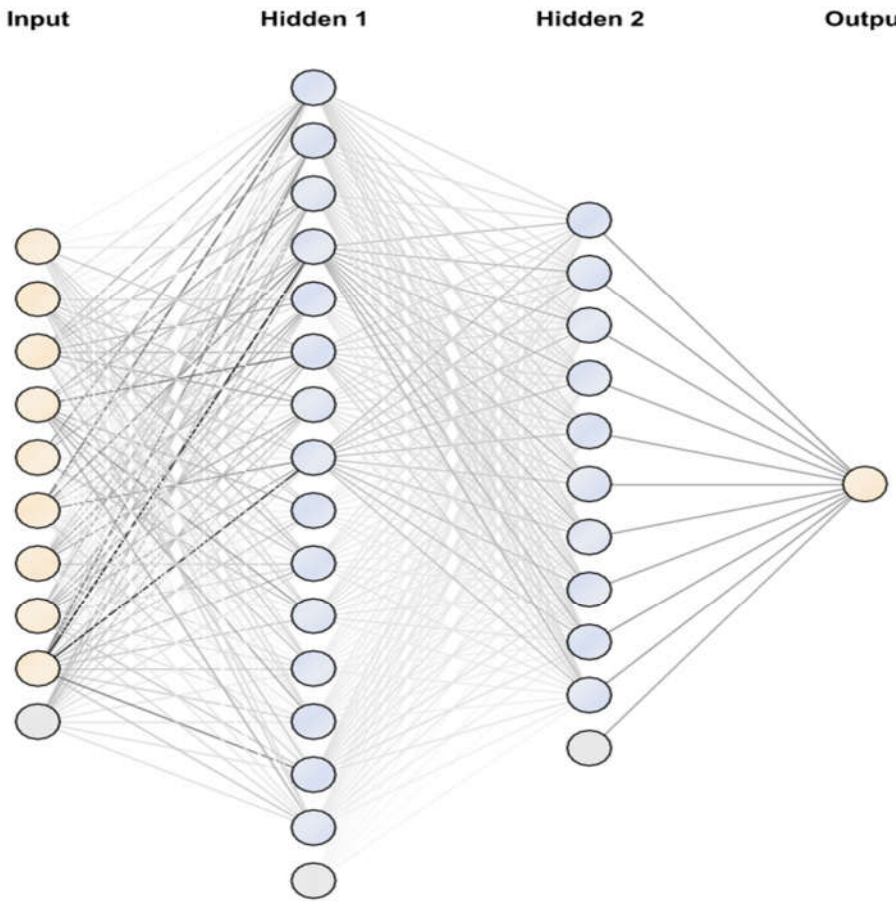

Figure 19: Neural Network Architecture for Final Cost Prediction

Table 11: weights for connection between input nodes (attributes) and hidden layer 1 nodes

Input Node	Hidden Layer - 1 Node														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
BH	0.135	-0.108	-0.145	-0.244	-0.043	-0.702	-0.107	-0.102	-0.476	-0.552	-0.46	-0.357	-0.447	-0.112	-0.439
BST	0.402	0.917	0.139	0.034	0.609	0.42	0.806	-0.454	0.582	0.846	0.574	0.637	0.438	0.126	0.552
DS	-0.915	0.028	0.856	0.358	-1.068	0.99	-1.135	0.629	0.054	-0.408	-0.122	-0.281	0.274	0.676	-0.032
ED	0.4	0.206	-1.211	0.907	-0.403	-1.851	-0.442	0.219	-0.965	-0.796	-1.286	-0.732	-1.075	-0.164	-0.977
FT	-1.347	0.708	0.354	-0.929	0.005	0.088	0.131	-0.283	-0.096	-0.081	-0.783	-0.078	-0.023	0.068	0
GNR	-1.995	0.283	-0.115	1.528	-0.544	-0.74	-0.389	1.452	-0.184	-0.23	0.283	-0.139	-0.294	0.568	-0.309
ST	0.195	-0.936	-0.82	-0.709	1.138	-0.707	1.148	0.131	0.263	0.734	0.258	0.446	0.041	-0.032	0.281
SW	0.557	-1.781	0.276	-0.292	0.993	1.005	1.013	-1.024	0.312	0.581	-0.104	0.442	0.416	-0.713	0.482
TSA	-0.581	-0.247	-0.87	-3.206	-0.576	-0.081	-0.773	-2.861	-0.731	-0.926	-0.778	-0.748	-0.666	-1.694	-0.733
Bias	-0.08	-0.108	-0.973	-0.85	-0.475	-0.793	-0.537	-0.459	-0.31	-0.4	0.229	-0.346	-0.42	-0.426	-0.425

Table 12: Weights for connections between nodes in hidden layer 1 and hidden layer 2

Hidden layer 1 Nodes	Hidden Layer 2 Nodes									
	1	2	3	4	5	6	7	8	9	10
1	-0.506	-0.428	-0.47	-0.518	-0.528	-0.485	-0.494	-0.432	-0.469	-0.493
2	-0.446	-0.425	-0.457	-0.399	-0.404	-0.416	-0.429	-0.431	-0.465	-0.479
3	-0.392	-0.378	-0.408	-0.435	-0.447	-0.364	-0.433	-0.409	-0.423	-0.407
4	-0.911	-0.939	-0.964	-0.958	-0.923	-0.97	-0.971	-0.918	-0.982	-0.928
5	-0.305	-0.341	-0.342	-0.33	-0.352	-0.31	-0.367	-0.317	-0.306	-0.329
6	-0.568	-0.574	-0.515	-0.541	-0.558	-0.569	-0.533	-0.545	-0.547	-0.541
7	-0.353	-0.38	-0.336	-0.363	-0.359	-0.39	-0.408	-0.345	-0.364	-0.366
8	-0.882	-0.819	-0.836	-0.84	-0.875	-0.789	-0.77	-0.883	-0.835	-0.764
9	-0.178	-0.216	-0.225	-0.207	-0.204	-0.222	-0.208	-0.25	-0.274	-0.236
10	-0.305	-0.347	-0.301	-0.266	-0.269	-0.248	-0.254	-0.287	-0.267	-0.312
11	-0.302	-0.376	-0.334	-0.299	-0.324	-0.358	-0.319	-0.332	-0.357	-0.258
12	-0.208	-0.17	-0.229	-0.173	-0.172	-0.212	-0.185	-0.199	-0.209	-0.25
13	-0.269	-0.214	-0.256	-0.283	-0.199	-0.259	-0.207	-0.233	-0.224	-0.199
14	-0.4	-0.338	-0.347	-0.37	-0.33	-0.382	-0.327	-0.387	-0.397	-0.34
15	-0.236	-0.211	-0.266	-0.223	-0.205	-0.255	-0.234	-0.274	-0.23	-0.194
Bias	-0.063	-0.102	-0.071	-0.097	-0.089	-0.086	-0.094	-0.106	-0.051	-0.126

5.1.4. GRADIENT BOOSTED TREES

The fourth technique used to develop a final cost estimation model is Gradient boosted trees. As discussed in the literature review GBT is an ensemble of decision tree models combined through use of boosting mechanism.

The number of trees (decision trees) and learning rate were determined through grid optimization whereas the remaining parameters (sample rate, minimum number of rows and the maximum depth for each decision tree) were identified through trial and errors. So, for every value set manually, the number of tree and learning rate was determined through grid optimization.

The GBT model was made of 388 decision trees and with a learning rate of 0.02. The model's performance was found to lack when compared to that of the neural network model with a MAPE of 37.22% +/- 21.83% and Absolute Error of 17321448.912 +/- 14339686.486. Whereas, compared to the decision tree model, it shows a slight improvement in performance. Table 13 provides the general description of the GBT model. The min, max and mean depth and leaves correspond to the tree depth and number of leaves in the base learner decision trees.

Table 13: General description of the GBT model for final cost prediction

No. of Trees	Min. Depth	Max. Depth	Mean Depth	Min. Leaves	Max. Leaves	Mean Leaves
388	7	10	9.28093	27	38	33.49742

The scoring history of the decision trees inside it is presented in Table 14. It can be seen that the training error has dropped with the number of trees. It shall be pointed that even though the relative error of the GBT model is slightly lower than the ANN model, the variance is much higher. The Absolute error, too, show that the GBT model performs lower than artificial neural network when it comes to final cost prediction.

Table 14: Scoring error of the GBT model with respect to the number of decision trees

Scoring History:

Timestamp	Duration	Number of Trees	Training RMSE	Training MAE	Training Deviance
2021-01-30 12:00:40	0.000 sec	0	50725216.93105	31258292.63216	5607.03153
2021-01-30 12:00:40	0.063 sec	1	49963152.78949	30502081.89473	5374.03001
2021-01-30 12:00:40	0.078 sec	2	49216527.59460	29753926.52458	5147.74829
2021-01-30 12:00:40	0.094 sec	3	48513911.60934	29062553.64214	4937.14688
2021-01-30 12:00:40	0.109 sec	4	47827383.18882	28393732.07475	4733.59001
2021-01-30 12:00:40	0.141 sec	5	47158940.43453	27742955.82543	4538.24795
2021-01-30 12:00:40	0.156 sec	6	46495609.20896	27083987.75798	4347.30807
2021-01-30 12:00:40	0.172 sec	7	45855405.84934	26467727.60846	4165.68005
2021-01-30 12:00:40	0.203 sec	8	45248838.89881	25892920.42542	3999.65086
2021-01-30 12:00:40	0.219 sec	9	44638868.40676	25327533.28358	3833.97049

2021-01-30 12:00:41	1.078 sec	40	29691034.64057	13821705.61162	1099.17152
2021-01-30 12:00:41	1.203 sec	41	29282108.68630	13581949.49191	1057.16954
2021-01-30 12:00:41	1.562 sec	42	28860188.35067	13329470.45883	1015.43155
2021-01-30 12:00:42	1.844 sec	43	28441579.54124	13091977.62167	974.76691
2021-01-30 12:00:42	2.098 sec	44	28025923.13742	12849412.86102	936.41030
2021-01-30 12:00:42	2.395 sec	45	27608667.49817	12590471.43627	897.01299
2021-01-30 12:00:42	2.663 sec	46	27203407.69088	12367621.45690	863.01502
2021-01-30 12:00:43	2.913 sec	47	26799972.95689	12141046.44158	829.22947
2021-01-30 12:00:43	3.147 sec	48	26399000.48834	11914024.33449	796.19429
2021-01-30 12:00:47	7.137 sec	388	14093293.60440	3303923.19730	86.18164

5.1.5. SUMMARY ON FINAL COST ESTIMATION

In the previous section, four modeling techniques were implemented to build a prediction model that predicts the final cost of building projects based historical final cost data of buildings built in Addis Ababa. As it can be seen from the results, the artificial neural network model has come out superior by producing the least amount of error with a relative error of 37.05%.

This is comparable to the Result of Smita K. and Adamuthe (2017) who developed NN models for Cost estimation of building projects and the NN model has an error of 42% and Tadesse and Dinku (2007) who used 48 projects' data to develop NN models for the prediction of Road projects in Ethiopia with an error of 32.58%.

Regardless, the performance gained even from the most accurate model, though it is within range for a class 4 estimate, when comparing this result with researches like El-Sawalhi (2015) who used 169 data and got error less than 6%, it is clear that this model is lagging behind. There could be different reasons for the weak performance of the model, ranging from the number of data instances the models were trained on to the quality of the data.

The performance of these models could improve with additional data. This is evident as those researches that used data set more than 100 instances have on average an error less than 10% where as those researches done with < 50 data sets has an average error of 22% (Patil & Salunkhe, 2020; El-Sawalhi, 2015; Bayram et al., 2013; Arafa and Alqedra, 2011; Elmousalami, 2019; Yadav et al, 2016).

This can be an indication that the humble performance of the ANN model in this research is a good indicator of the potential of ANN in Predicting the cost of building projects, provided that enough data is used to train the model. But considering the challenges faced in collecting such data, especially in countries like Ethiopia where most data are not managed and organized in a database, this might not be an ideal solution.

Furthermore, other researchers have managed to come up with better performances with data as small as, and in some cases, smaller than 39 data points. Case in point is the study by Bhirud & Ambrule (2017) who developed Final cost prediction model using ANN and only 12 buildings' data and managed to develop an ANN model with an average error of 16.1% another one is the study of Ji et al. (2011).

Another reason for such poor performance can be the quality of the data, particularly with regards to the models (in) ability to account for the fluctuation of cost within the data even after adjusting for inflation using Consumer Price Indices. A report by the Canadian construction association (2012) identified using non-industry specific cost indices as among the reasons for poor performance of estimation models in the construction industry – a notion supported by Zhang (2017), who, after accounting for economic fluctuations as dependent variables in his models, concluded that these models will not work in environments with high economic fluctuations.

Owing to that, the next part focuses on an alternative approach – the Quantity – based approach by comparing the cost – based and quantity –based approach for the case of structural cost estimation.

5.2. STRUCTURAL COST PREDICTION

The second main objective of this study is to make a comparison between the cost – based and quantity – based approaches for the case of Structural cost prediction. In order to do that first a structural cost prediction model is built using the Cost-based approach. Following that, prediction models will be built for the quantity of works that make up the main structural work of a reinforced concrete building, namely Concrete work, Formwork and reinforcement work.

5.2.1. COST- BASED APPROACH

In this section the performance of Linear Regression, Neural Network, Decision Tree and Gradient boosted Trees will be evaluated for the case of structural cost prediction of a building.

4.3.1.1. DATA DESCRIPTION

Since the data used is similar to the one used for Final Cost modeling, only data pertaining to Structural cost is discussed here. The range of structural cost in the data is between 2,676,461 birr and 93,051,934 birr with average of 17,065,959. The frequency distribution of the structural cost is shown in Figure 20.

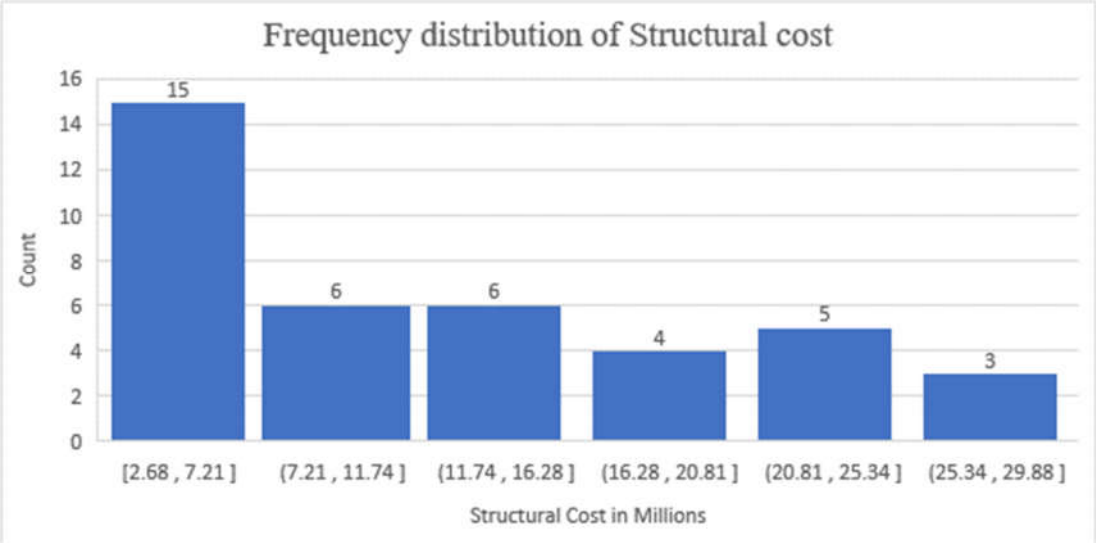

Figure 20: Frequency distribution of structural cost

4.3.1.2. VARIABLE SELECTION AND MODELING

The independent Variables to be used in developing the structural cost models are identified from literature review in section 2.4 and presented in Table 6. The modeling result is discussed as follows:

A. Linear Regression

The linear regression model is yet again the worst performing technique with a relative error of 43.57% +/- 32.27%. Actual Vs. predicted plot is shown in Figure 21, as it can be seen the predictions are far from the fit line.

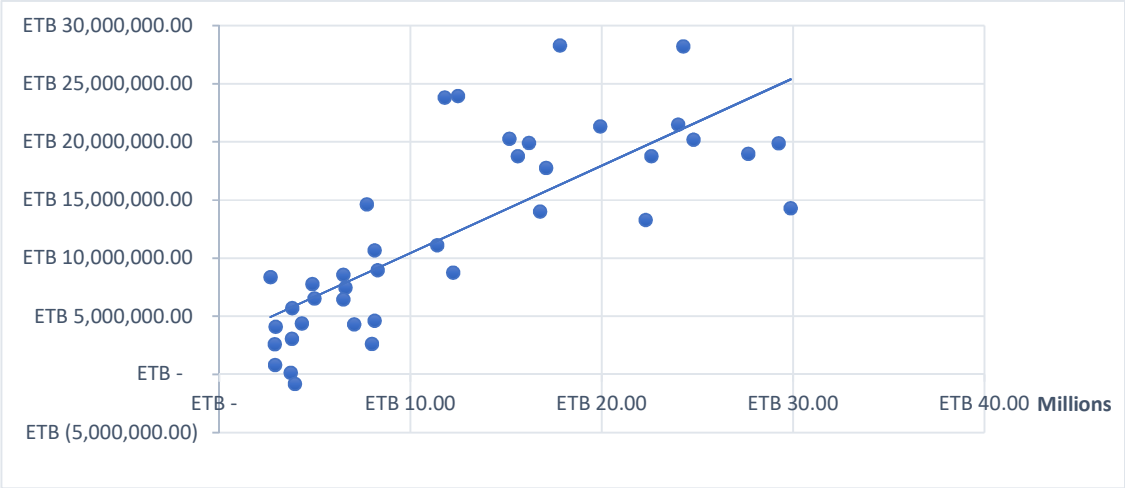

Figure 21: Actual vs. Predicted Structural Cost plot for Linear Regression

B. Neural Network

In order to come up with the best neural network architecture, a total of 8866 combination of parameters for learning rate, training cycles (epochs), momentum as well as the number and size of hidden layers were tried using a grid Parameter Optimization tool. The best neural network for structural cost prediction has 2 hidden layers each having 4 nodes. The learning rate, epoch and momentum were 0.108, 80 and 0.24 respectively.

The neural network had a rather high relative error of 34.71% +/- 19.48% and an absolute error of 3721260.404 +/- 1843808.674. This is even worse than the decision tree model.

C. DECISION TREE

The Decision tree model performed better than LR and ANN with a relative error of 27.29% +/- 8.74% having lower bias as well as variance error. Since the actual decision tree is too large to put on a paper, the rules to produce the regression tree is provided in Figure 22.

RegressionTree

```
Total Slab  > 2923.500
|   Total Slab  > 5425.500: 36161062.640 {count=2}
|   Total Slab  ≤ 5425.500
|   |   Average Floor Height > 2.975
|   |   |   Average Floor Height > 3.320
|   |   |   |   Slab thickness (typical floors > 16.500: 22442942.860
{count=2}
|   |   |   |   Slab thickness (typical floors ≤ 16.500: 16346714.440
{count=2}
|   |   |   Average Floor Height ≤ 3.320
|   |   |   |   Average Floor Height > 3.064: 24343891.227 {count=3}
|   |   |   |   Average Floor Height ≤ 3.064: 29562971.705 {count=2}
|   |   Average Floor Height ≤ 2.975
|   |   |   Slab type > 0.500: 17288957.615 {count=2}
|   |   |   Slab type ≤ 0.500: 12705798.270 {count=4}
Total Slab  ≤ 2923.500
|   Total Slab  > 1907.500
|   |   Total Slab  > 2471: 6818420.188 {count=2}
|   |   Total Slab  ≤ 2471: 14208063.600 {count=2}
|   Total Slab  ≤ 1907.500
|   |   No. of Floors  > 4.500
|   |   |   Foundation type > 0.500: 8194790.203 {count=2}
|   |   |   Foundation type ≤ 0.500: 8053766.240 {count=2}
|   |   No. of Floors  ≤ 4.500
|   |   |   Total Slab > 1130
|   |   |   |   Slab thickness (typical floors > 22
|   |   |   |   |   Average Floor Height > 2.750: 4916605.782 {count=2}
|   |   |   |   |   Average Floor Height ≤ 2.750: 3542748.779 {count=3}
|   |   |   |   Slab thickness (typical floors ≤ 22: 6894981.426 {count=3}
|   |   |   Total Slab ≤ 1130
|   |   |   |   Average Floor Height > 3.367: 4056386.666 {count=2}
|   |   |   |   Average Floor Height ≤ 3.367
|   |   |   |   |   No. of Floors > 2.500: 2784849.224 {count=2}
|   |   |   |   |   No. of Floors ≤ 2.500: 3351513.823 {count=2}
```

Figure 22: Decision/Regression Tree rules for Concrete Quantity

D. GRADIENT BOOSTED TREES

The gradient boosted trees had the best performance in predicting the structural cost in this study. The relative error and absolute error produced are 22.67% +/- 7.70% and 3043467.624 +/- 2416769.620. The best GBT model was built of 32 decision trees as a base learner. A summary of the model is provided in Table 15.

Table 15: General characteristics of GBT model for structural cost estimation

Number of Trees	Number of Internal Trees	Model Size in Bytes	Min. Depth	Max. Depth	Mean Depth	Min. Leaves	Max. Leaves	Mean Leaves
32	32	6315	5	5	5	7	16	11.09375

5.2.2. QUANTITY BASED APPROACH.

In this phase the cost and quantity-based approaches are evaluated by developing quantity of works estimation models for Concrete, Formwork and Reinforcement works to determine structural cost of buildings and comparing this approach with the accuracy of a cost-based structural cost prediction.

First all four modeling techniques are used to build models that predict the quantity of works for Concrete, Formwork and Reinforcement works. The result for this is provided as follows:

5.2.2.1. CONCRETE WORKS QUANTITY ESTIMATION

A. LINEAR REGRESSION

The performance of linear regression was evaluated for concrete quantity estimation. The performance of each regression model is presented in Figure 23. As it can be seen, the regression model doesn't seem to perform well on the concrete volume (quantity) estimation. The LR model has a relative error of 33 % and a coefficient of determination of almost 80%.

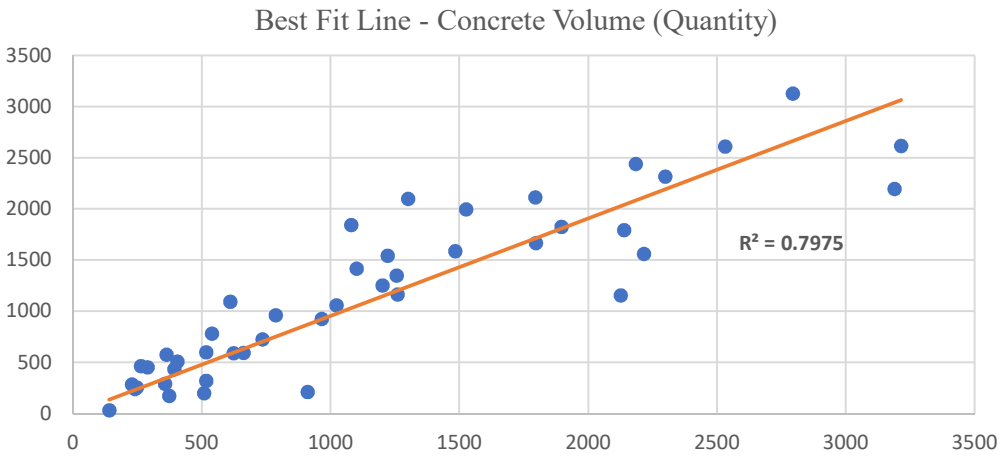

Figure 23: Best Fit Line for Concrete Quantity Estimation using Linear Regression

B. Decision Tree

The decision tree model does seem to perform slightly better than the linear regression model with a relative error of 27.27%. The decision tree for Concrete quantity estimation is provided in Figure 24.

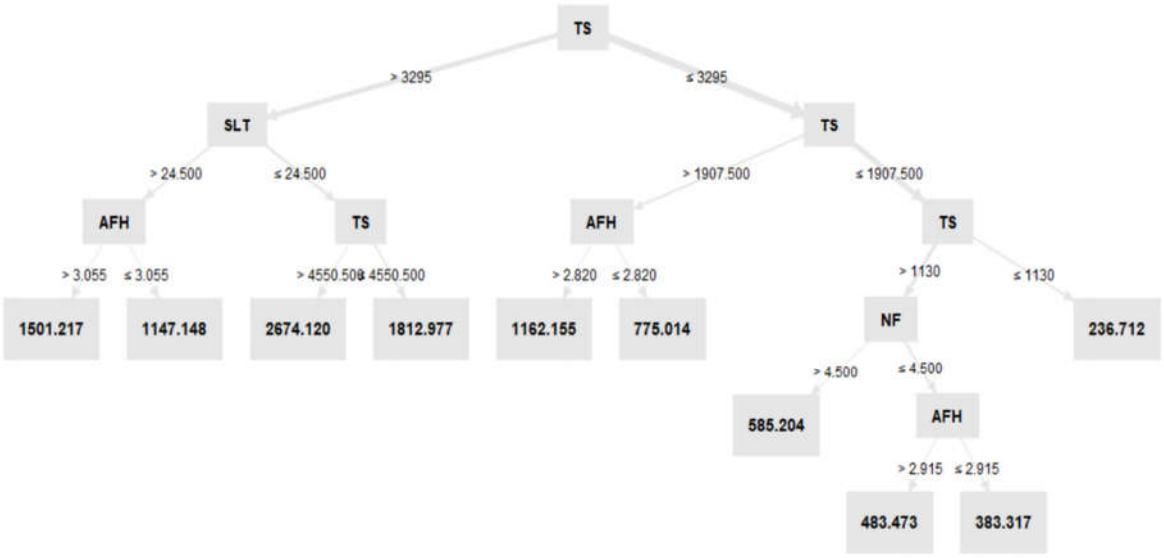

Figure 24: Decision Tree for Concrete Quantity Estimation

C. Neural Network

Perhaps a significant increase in performance is seen when NNs are adopted for concrete volume prediction, which resulted in a relative error of 16.44% +/- 4.91% and an absolute error of 168.438 +/- 109.279. This is a significant improvement from the linear regression and Decision tree models. In Figure 25, the actual vs. Predicted graph extracted from the 10-fold cross validation is presented. As it can be seen, the prediction is quite close to that of the actual values and follow similar trend.

The architecture for the best NN model for Concrete Quantity prediction is comprised of one hidden layer having 5 nodes. The learning rate, Training cycle and momentum for the best NN model were 0.008, 200 and 0.9 respectively. The weights for the nodes are presented in Table 16

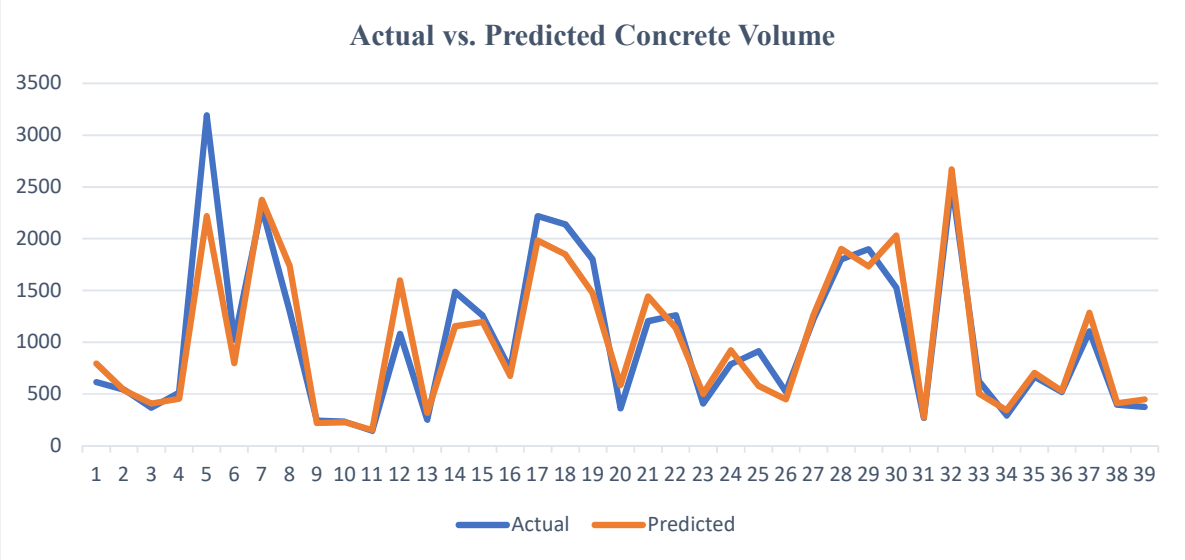

Figure 25: Actual vs. Predicted Concrete volume extracted from 10-fold cross validation result of a NN model

D. Gradient Boosted Trees

The GBT model's performance was found to be lagging behind the NN model. For the case of Concrete quantity prediction and came in third with an AE of 213.416 +/- 163.000 and Relative error of 20.22% +/- 11.70%

Table 16: Weights for connections between nodes in Input layer and first hidden layer (top) and between output node and nodes in the first hidden layer (bottom)

Input Layer	Hidden layer Nodes					
	Node 1	Node 2	Node 3	Node 4	Node 5	
AFH	0.051	-0.087	-0.691	-0.041	-0.126	
BST	-0.004	-0.072	-0.773	-0.076	-0.019	
FT	0.092	0.123	-0.711	0.15	0.203	
NF	-0.015	-0.154	0.188	-0.153	-0.081	
NL	-0.103	0.075	-0.171	0.023	0.34	
SLT	0.028	0.09	0.578	0.076	-0.048	
ST	-0.072	-0.041	-0.326	-0.023	0.23	
SW	-0.059	0.234	0.03	0.274	0.577	
TS	-0.11	-0.71	-1.851	-0.552	-1.099	
Bias	-0.056	-0.113	0.628	-0.112	-0.122	
Output Layer weight						
	Hidden layer nodes					
	Node 1	Node 2	Node 3	Node 4	Node 5	Threshold
Output node	0.268	-0.391	-1.74	-0.28	-0.925	1.488

5.2.2.2. REINFORCEMENT WORKS QUANTITY ESTIMATION

❖ <u>Data description with respect to Reinforcement cost and quantity</u>

The quantity of reinforcement bar used in the buildings range from 10,763 Kgs to 377,448 KGs with the majority of the data being below 230,774 kg. The frequency distribution for the parameter, Rebar quantity, is provided in *Figure 26*. The cost of reinforcement works ranges from six hundred ninety thousand to twenty-three and a half million. The percent share of the reinforcement cost in the total structural cost ranges from a minimum of 39% up to 85% with the median being 55%. 10-fold cross validation was used to validate the models and the performance of all four modeling techniques was assessed using the performance metrics identified in the methodology.

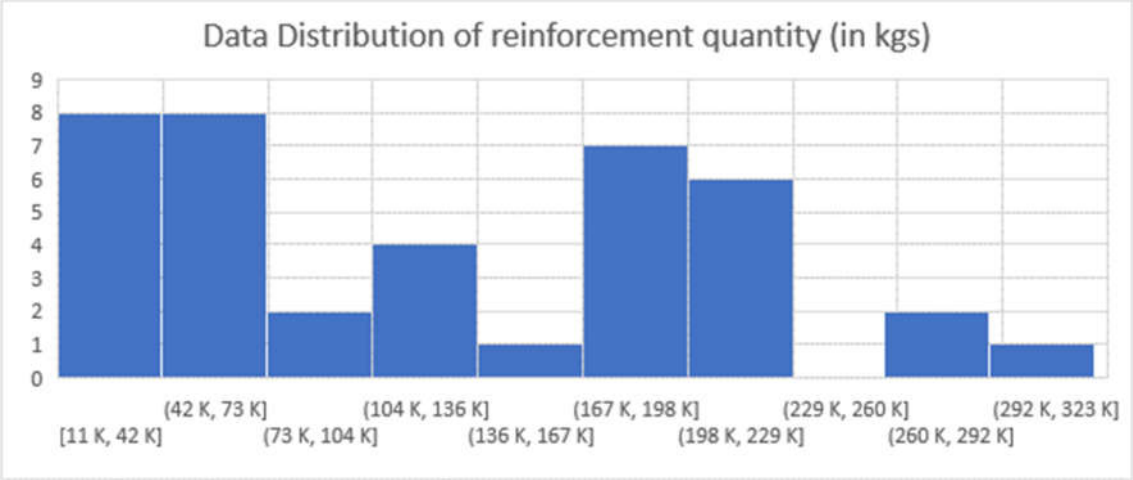

Figure 26: Frequency distribution of Rebar Quantity in buildings for the data in question

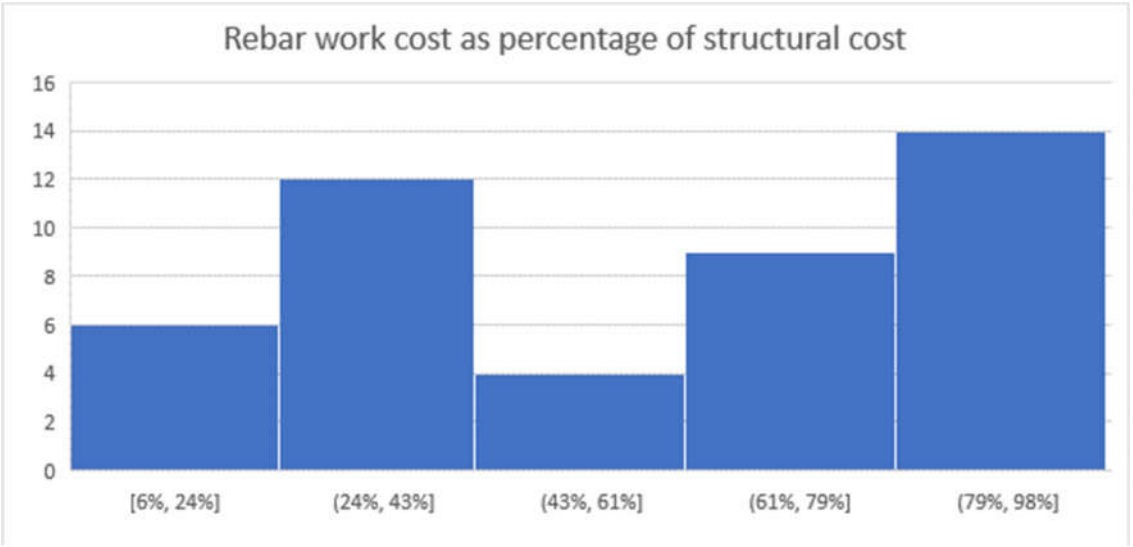

Figure 27: Frequency distribution of the Percent share of rebar cost in total structural cost for the data

A. Linear Regression

The performance of the Linear Regression for both Rebar quantity and Cost is presented in Table 17.

Table 17: Performance of Linear Regression in Estimating Rebar Quantity

Model type	Absolute Error	Relative Error
Rebar Quantity	28,233.990 +/- 8,483.805	38.61% +/- 43.78%

Furthermore, just like the case of Concrete quantity estimation, negative values of the dependent variables were recorded when performing cross validation on the models. Based on

the disappointing performance, it is safe to say there seem to be no sound linear relationship between the independent variables and the label variable.

B. DECISION TREE

The performance for the decision tree for Rebar quantity prediction is tabulated in Table 18. The Decision tree model seem to perform less than the Linear Regression model in estimating the quantity of Reinforcement works but on a relative scale, the DT model have lower relative errors. The decision trees developed for is presented in Figure 28.

Table 18: Performance of Decision tree in estimating the quantity of reinforcement works

Model type	Absolute Error	Relative Error
Rebar Quantity	30,042.388 +/- 13,094.794	29.78% +/- 9.54%

C. NEURAL NETWORK

The neural network model has performed much better than the linear regression and Decision tree models for Rebar quantity estimations. The best neural network model for Rebar quantity estimation has one hidden layer with only one node. The performance for the NN models is summarized in `Table 19`.

Table 19: Performance of NN in predicting the Quantity of Reinforcement Works

Model type	Absolute Error	Relative Error
Rebar Quantity	19276.874 +/- 10478.414	19.32% +/- 7.80%

Based on the relative errors, the NN model seem to perform much better when applied to the rebar quantity problem as compared to the cost estimation one. The weights for each node are summarized in Table 20.

Table 20: Weights for connections between Input layers (Attributes) and Hidden Layer 1 nodes (top) and between 1st hidden layer nodes and the output node – Rebar Quantity

	Rebar Quantity Model	
	Hidden Layer 1	
Input Layer	Node 1	Threshold
Slab type	-0.126	
Foundation Type	-0.173	
Basement floor	0.209	
Number of floors	-0.827	
Slab thickness	-0.175	
Number of lifts	0.141	
Average floor height	-0.759	
Shoring work	-0.017	
Total slab area	-1.666	
Bias	-0.197	
Output Layer		
Hidden Layer nodes	Node 1	
Output layer		
	-2.09	0.904

D. GRADIENT BOOSTED TREES

The gradient boosted tree model for the reinforcement Quantity was built using 200 decision trees as base learners. The NN model was found to be superior in estimating rebar quantity. The general information on the GBT models is presented in Table 21.

Table 21: General Description and performance of GBT models for Rebar cost and Quantity estimation

Model	No. of Trees	Min. Depth	Max. Depth	Min. Leaves	Max. Leaves	Absolute Error	Relative Error
GBT Rebar Quantity	200	4	9	9	12	24642.454 +/- 9586.763	23.79% +/- 10.70%

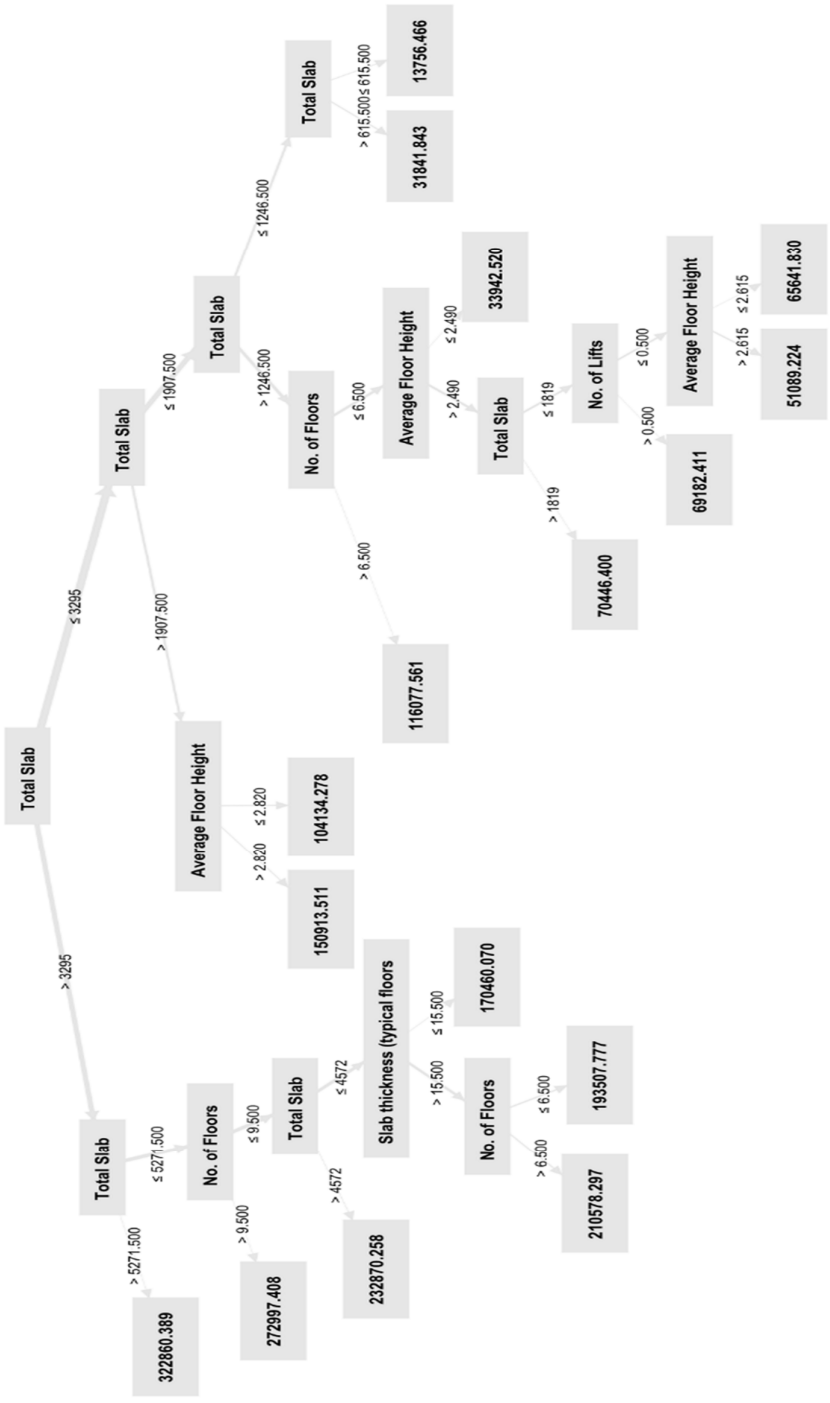

Figure 28: Decision tree for prediction of Reinforcement work Quantity

5.2.2.3. FORMWORK COST AND QUANTITY ESTIMATION

Description of Data With regards to Formwork quantity

The formwork quantity comprises of the total area of formwork used for the construction of structural members (both suspended and unsuspended) of the buildings. The value for the formwork quantity ranges from 626 m^2 to 20,928 m^2. The formwork cost of the buildings', on the other hand, ranges from 119,253 birrs to 7,231,563 birr and covers between 1.47% and 24.1% of the total structural cost, with an average of 15.85%.

A. LINEAR REGRESSION

The performance and equations for the linear regression models is provided in Table 22. As it can be seen, the formwork cost models seem to perform very badly whereas significant improvement is seen for the formwork quantity model.

Table 22: Performance and equation of Linear Regression model for estimating formwork quantity

Model	Abs. Error	Rel. Error	Equation
Formwork Quantity	1362.954 +/- 595.066	25.92% +/- 13.68%	1766.305 * ST + 763.900 * FT + 2130.645 * BST + 42.952 * NF + 40.397 * SLT – 30.821 * NL – 375.584 * AFH – 384.969 * SW + 2.398 * TSA – 1320.394

- Based on the correlation matrix shown in Figure 29, the total slab area (TSA) has a high correlation with the formwork quantity.

B. DECISION TREE

The performance of the decision tree models is summarized in Table 23. The performance of DT for the case of Formwork quantity estimation was found to be superior to the linear regression model but the Absolute error of the LR model is slightly better.

Table 23: Performance of Decision Tree in estimating the Cost and quantity of formwork

Model type	Absolute Error	Relative Error	RMSE	Error squared
Formwork Quantity	1385.910 +/- 949.994	21.57% +/- 12.44%	1874.257 +/- 1509.593	5563 X 10^3 +/- 8377 X10^3

C. Neural Network

Based on the previous track record of NNs, one would assume they will provide more accurate results for this case as well but on a first glance, the NN seem to perform slightly worse than Decision trees for both formwork quantity estimation. It can be seen from Table 23 and Table 24, the NN model have an absolute error that are lower than that of the DT but the relative error is higher than the DT model.

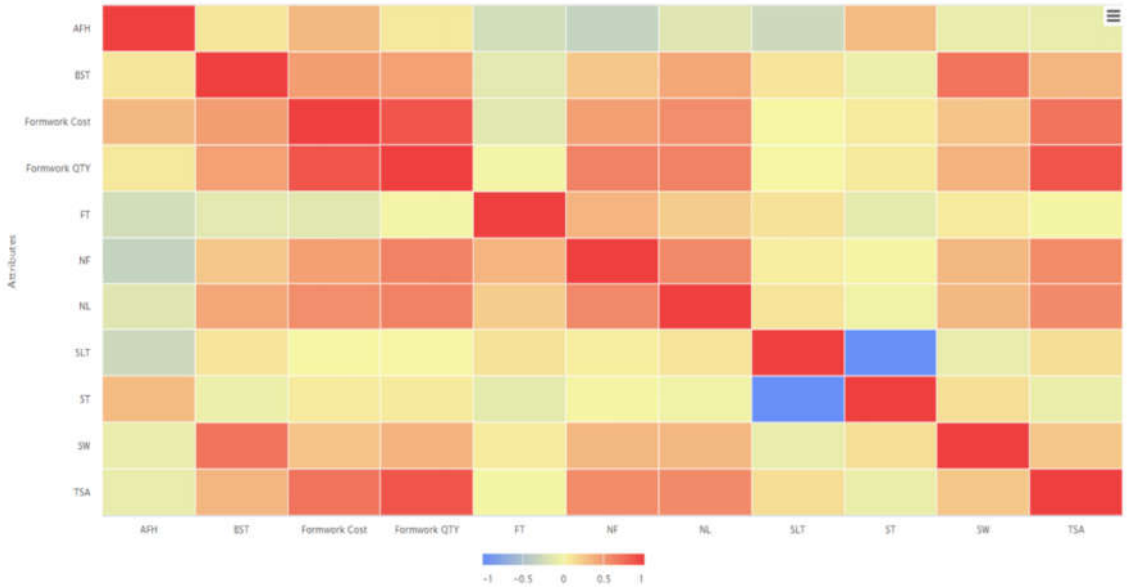

Figure 29: Correlation matrix between attributes and formwork area (Qty)

This means, depending on which performance metric we are using, we can have two different winners for the same problem. For the case of the formwork cost, if we are using the Relative error, the decision Tree will come out as a winner whereas if we are using the Absolute error as a measure then the NN model will be the better option. The reason for this as noted by hssay (2018) is the wider range of the output variable – that is the wider the range is for the predicted variable; different relative error could be gained for the same absolute error. In order to solve this tie between them, two more performance metrics, the root mean square error and the squared error, were adopted. Based on these results, it can be seen that The NN model has a higher performance as compared to the Decision tree and linear regression models for the case of Formwork quantity estimation.

Table 24: Performance of NN in estimating the quantity of formwork

Model type	Absolute Error	Relative Error	RMSE	Squared Error
Formwork Quantity	1349.236 +/- 751.458	24.54% +/- 17.53%	1795.264 +/- 1160.790	4435662.719 +/- 5656346.072

D. GRADIENT BOOSTED TREES

The absolute and relative error for the GBT models is provided in Table 25. The GBT model showed superior performances in estimating formwork quantity as they have the least errors in both performance metrics. The GBT model used for the formwork cost estimation was built using 150 decision trees. The general description of the models is provided in Table 26.

Table 25: Performance of GBT in estimating Formwork Quantity

Model type	Absolute Error	Relative Error
Formwork Quantity	1119.994 +/- 281.695	19.58% +/- 9.52%

Table 26: General Description of GBT model adopted for Formwork cost and quantity prediction

Prediction	Number of Trees	Min. Depth	Max. Depth	Min. Leaves	Max. Leaves	Shrinkage (learning) rate
Formwork Quantity	150	4	8	7	11	0.05

5.2.3. COMPARISON BETWEEN THE COST BASED AND QUANTITY BASED APPROACHES

In the previous section, four modeling techniques were used to develop prediction models for quantity of Concrete, Reinforcement bar and Formwork. In order to compare the quantity-based approach with the cost based one for the case of structural cost of building projects, The Structural cost was also first determined using the cost – based method.. Table 27 presents a summary of the winning model's performance for each prediction. As it can be seen, the Neural network and Gradient boosted trees performance is comparable.

It can also be seen from the table that the absolute errors for Cost and quantity predictions are not on the same scale and taking the absolute error of each models directly for comparison will not provide a fair evaluation. Thus, it was proposed to convert the absolute errors for the quantity of works into cost by using an average unit rate for the year 2019. Based on the collected data, the minimum, average and maximum unit rate prices for each work items are provided in Table 28.

Table 27: Summary of performance of Wining Models for different cost and quantity estimations

Approach	Item	AE	RE	Modeling technique
Quantity Based	Concrete	168.438 +/- 109.279	16.44% +/- 4.91%	NN
	Rebar	19276.874 +/- 10478.414	19.32% +/- 7.80%	NN
	Formwork	1119.994 +/- 281.695	19.58% +/- 9.52%	GBT
Cost based	Structural cost	3043467.624 +/- 2416769.620	22.67% +/- 7.70%	GBT

Table 28: minimum, maximum and average unit rates for work items concrete, reinforcement and formwork. Rates Collected from buildings built in 2019.

Work Item	Unit Rate		
	min	average	Max
Concrete	2,398.969	3,177.418	4,430.117
Reinforcement	40	52.88451	70
Formwork	279	317.3563	350

To compare which approach produces more accurate results based on their absolute errors, the absolute error for the quantity models is multiplied by the unit rates set out in Table 28. This is equivalent to multiplying the actual and predicted values of the quantity of works and multiplying them by the unit rates and then calculating the following table summarizes the calculation.

Table 29: Absolute errors of best models (multiplied by unit rates) for concrete, formowk and rebar

Work item	Absolute error (multiplied by unit rate)		
	min	**average**	**max**
Concrete	404077.540422 +/- 262156.9	535197.933084 +/- 347225.0616	746200.047246 +/- 484118.7556
Rebar	771074.96 +/- 419136.6	1019448.03582174 +/- 554145.79	1349381.18 +/- 733488.98
Formwork	312478.326 +/- 78592.91	355437.1518622 +/- 89397.68293	391997.9 +/- 98593.25

Now that the absolute errors are scaled, the next problem before making comparison is the matter of error propagation.

The Case of Error-propagation

The whole essence of the quantity-based approach is to predict the quantity of works first and then to estimate their cost using unit rates for the respective work items - thereby reducing the impact of cost fluctuation and inflation that affects a cost data. In the previous section, models were developed to predict the quantity of works that make up the structural work of a reinforced concrete building, namely; concrete, reinforcement and formwork. Using the quantity-based approach, one can determine the total structural cost by first multiplying these quantity predictions with their respective unit rate to get the cost of individual work item and summing these costs.

It is to be noted that these costs are not known in certain as they themselves are a function of a unit rate and a quantity of works that was determined from a prediction of a model for each respective work items and thus, hold some uncertainty in them. The question then becomes, how can one be sure that the structural cost that is generated by summing the product of a predicted value for the quantity of work items and a unit rate price is more accurate than just predicting the structural cost altogether, especially considering the propagation of uncertainty that arises from adding uncertain values (the concrete, rebar and formwork costs) to get a final

value (structural costs). To answer this question, the 'propagated' absolute error for the quantity-based approach is first calculated.

Statistically speaking, the error of a function that is made of uncertain variables will depend on the error of the independent variables. This is what is called Propagation of Error or Uncertainty. The Propagated error of the function will depend on the mathematical operator used to combine the independent variables. For this case, the mathematical operator adopted is addition as the costs of concrete, Rebar and Formwork are added to calculate the structural cost. Accordingly, the propagated absolute error for the Structural cost, εS, is given by:

$$\varepsilon S = \sqrt{(\varepsilon C)^2 + (\varepsilon R)^2 + (\varepsilon F)^2} \qquad (15)$$

Where; $\varepsilon C, \varepsilon R$ & εF are Absolute errors for Cost estimation of Concrete, Reinforcement and Formwork respectively (Vern, 2000).

These absolute errors for each work item were calculated in Table 29 for the minimum, maximum and average unit rate values. Taking the maximum absolute values, the propagated error was calculated as follows:

$$\varepsilon S = \sqrt{746200.047246^2 + 1349381.18^2 + 391997.9^{\,2}}$$

$$\varepsilon S = \sqrt{2{,}531{,}306{,}433{,}052.54} = 1{,}591{,}007.993$$

Similarly, for the standard deviation of the absolute error:

$$\varepsilon S = \sqrt{484118.7556^2 + 733488.98^2 + 98593.25^{\,2}}$$

$$\varepsilon S = \sqrt{782{,}097{,}682{,}292.33} = 884{,}362.868$$

The propagated absolute error is then: **$\underline{1{,}591{,}007.993 \pm 884{,}36\ 286\ 8}$**this is the actual error of the quantity based approach for structural cost estimation.

Comparing the Absolute error of the cost-based structural cost estimation and the propagated absolute error of the quantity-based structural cost estimation, it is clear that the quantity-based approach produce fewer errors.

Table 30 summarizes the different prediction models built and the performance of the winning modeling techniques.

Table 30: Relative and Absolute error of best models for each predictions made – a summary

Prediction	Best technique	Relative Error	Absolute Error	Propagated error
Final Cost	NN	39.43%+/- 10.08%	16632001.520+/- 9481894.959	
Concrete Quantity (volume)	NN	16.44% +/- 4.91%	168.438 +/- 109.279	1,591,007.993 +/- 884,362.868
Rebar quantity	GBT	19.32% +/- 7.80%	19276.874 +/- 10478.414	
Formwork Quantity	GBT	19.58% +/- 9.52%	1119.994 +/- 281.695	
Structural Cost	GBT	22.67% +/- 7.70%	3043467.624 +/- 2416769.620	

5.3. VARIABLE IMPORTANCE

Once the models were built, validated and the best techniques identified, the importance or weight of the independent variables in predicting the costs and quantities was extracted. This is especially necessary to identify what kind of relationship exist between the independent and dependent variable.

The importance of a variable for the NN models was identified by using the LIME algorithm which outputs a list of explanations on the contribution of each attribute to a prediction on the data sets. The algorithm first generates samples closer to each example and then a correlation weight is calculated the attributes and the output. Based on this correlation weight, attributes with positive weight are considered as important for the prediction while those with negative correlation weight are considered less important for the prediction. This will be carried out for each data points. Since the importance of the variables is determined by the correlation weights rather than the actual weights of the variables in the NN model, the sum of the weights is not equal to one.

5.3.1. FINAL COST PREDICTION

The correlation weights for each attribute used in the model is provided in Figure 30 . The total slab area (TSA) has the largest weight in this neural network model and thus has the highest importance whereas Slab type (ST) and Basement (BST) has the least importance. This could perhaps be the reason why practitioners use a price per square meter of the building gross area to estimate the final cost of a project at the preliminary stage. Even then, TSA, though it has a large importance in estimating a final cost of a project, it alone cannot explain for all predictions

without the help of the remaining attributes like External Decoration (ED), Foundation Type (FT), etc.

5.3.2. STRUCTURAL COST ESTIMATION

- ✓ Best Model = GBT with 32 Decision trees as a base Learner.

One major advantage of Gradient boosted trees over Neural Networks is the interpretability of the models. The variable importance in a GBT model is the average sum of the importance of the variable in the base learner and can be calculated using the equation:

$$I_l^2 = \frac{1}{M}\sum_{m=1}^{M} I_t^2(T_m) \qquad (16)$$

Where I_l^2 is the squared relevance for each attribute x_l in a single decision tree and is given by:

$$I_l^2(T) = \sum_{t=1}^{J-1} I_t^2 I(v(t) = l) \qquad (17)$$

Where $I_t^2 I(v(t) = l)$ is the sum of the squared improvements over all internal nodes for which the attribute was chosen as a splitting variable. (Trevor Hastie, 2017)

The Total slab area has, yet again, the highest importance with the AFH coming in second. The number of lifts and shear wall has the least importance in estimating the structural cost of building projects. The relative and scaled importance of the attributes in Structural cost estimation is provided in Table 31.

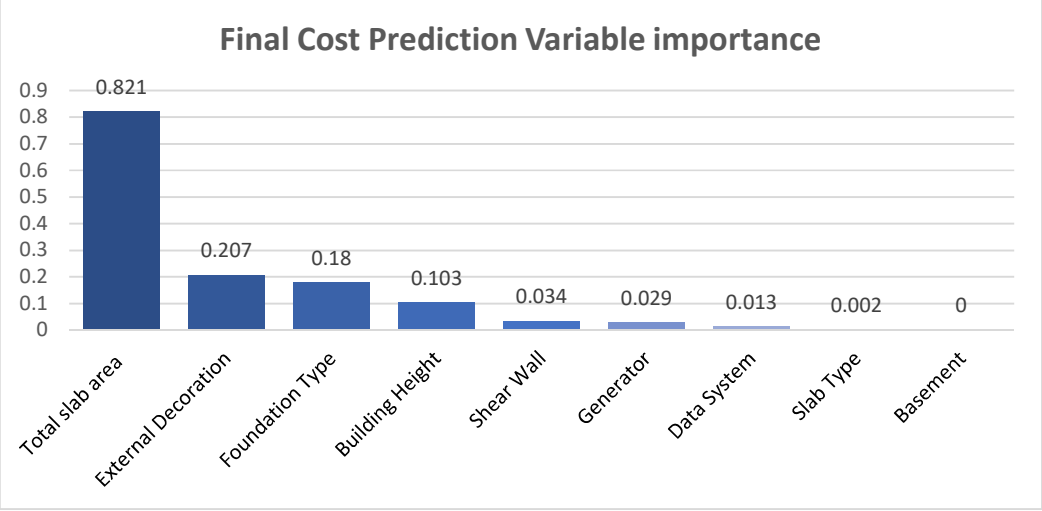

Figure 30: Importance (correlation) weight of Attributes for Final cost prediction

Table 31: Variable importance in Estimating Structural Cost

Variable	Relative Importance	Scaled Importance	Percentage
Total Slab Area	227744096	1	0.827923
Average Floor Height	39559108	0.1737	0.14381
Number of floors	3732289.5	0.016388	0.013568
Slab type	3348267	0.014702	0.012172
Foundation type	288411.125	0.001266	0.001048
Slab thickness	205009.2656	0.0009	0.000745
Basement	161563.8125	0.000709	0.000587
Shoring work	22236.2207	0.000098	0.000081
Number of lifts	17919.58984	0.000079	0.000065

5.3.3. CONCRETE VOLUME (QUANTITY) ESTIMATION

- ✓ Best Model = Neural network with One hidden layer having 5 nodes.

The attributes that have the highest importance in predicting the concrete volume are the Total slab area (TS), Basement Levels (BST) and Average Floor Height (AFH) whereas Shoring work (SW) and Slab thickness (SLT) have the least importance, with Slab thickness having 0 importance. The weights for all attributes used is shown in Figure 31.

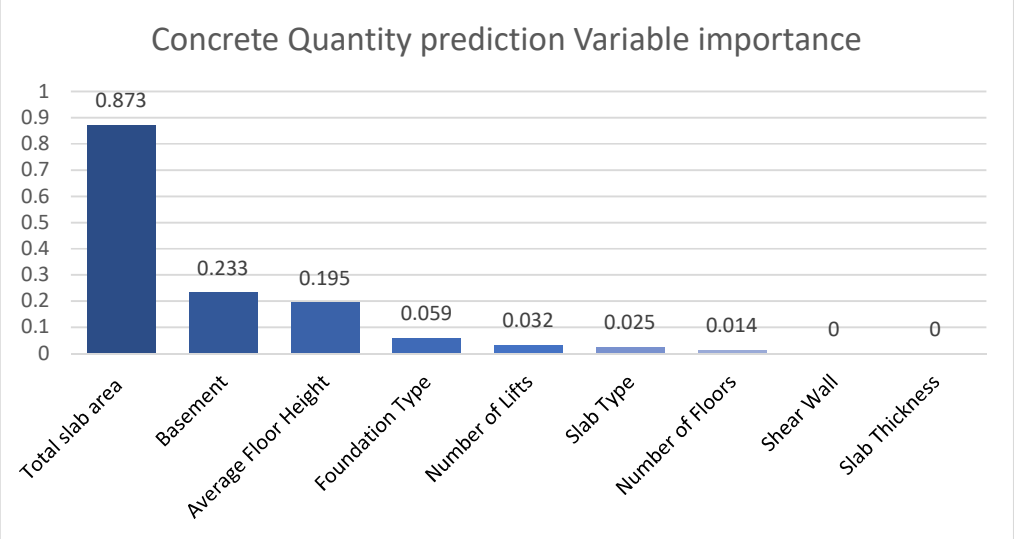

Figure 31: Importance (correlation) weight of Attributes for Concrete Quantity prediction

5.3.4. REINFORCEMENT QUANTITY ESTIMATION

✓ Best Model = Gradient Boosted tree with 200 decision trees as a base learner. The variable importance for the prediction of reinforcement quantity is summarized in Table 32.

Table 32: Variable importance for Prediction of Reinforcement Quantity

Variable	Relative Importance	Scaled Importance	Percentage
Total Slab Area (TSA)	1.27427E+12	1	0.90402
Average Floor Height (AFH)	59825381376	0.046949	0.042443
No. of Floors (NF)	42397200384	0.033272	0.030078
Foundation type (FT)	11005288448	0.008637	0.007808
No. of Lifts (NL)	9260142592	0.007267	0.00657
Slab thickness (SLT)	5624888320	0.004414	0.003991
Slab type (ST)	4003256832	0.003142	0.00284
Basement (BST)	2274891264	0.001785	0.001614
Shoring Work (SW)	897882304	0.000705	0.000637

5.3.5. FORMWORK QUANTITY ESTIMATION

✓ Best models = GBT with 150 decision trees as a base learner

The importance of the attributes is summarized in Table 33. As it can be seen, total slab area still has the highest importance followed by the number of floor.

Table 33: Variable importance in Estimating Formwork Quantity

Attribute	Relative Importance	Scaled Importance	Percentage
Total Slab Area (TSA)	7638927872	1	0.869259
No. of Floors (NF)	402883904	0.052741	0.045846
Basement (BST)	224249696	0.029356	0.025518
Average Floor Height (AFH)	223359296	0.02924	0.025417
Shoring works (SW)	135961184	0.017798	0.015471
Slab thickness (SLT)	57812984	0.007568	0.006579
Number of lifts (NL)	49121372	0.00643	0.00559
Foundation type (FT)	32074172	0.004199	0.00365
Slab type (ST)	23467066	0.003072	0.00267

5.4. MODEL DEPLOYMENT

The ultimate goal of building a predictive model is to use it in production to estimate or predict a certain variable. This is called model deployment. There are different ways one could deploy a model. One can create an excel sheet with the independent variables leaving the dependent variable empty, insert it to the data mining software as a test set and run the model to predict the dependent variable. This requires the user to be familiar with the data mining algorithm but, mostly, those in the management position are hardly familiar with such software and thus is not an ideal approach.

The other option is to export the model built and import it to a standalone program custom built for predicting the output variable. The users will be able to manually input the values for the independent variables and the program will process the input variables, call the trained model to make the prediction and output the predicted value to the user. Making a standalone program from the ground up will require a programming skill and a significant amount of time and effort.

To save time and effort, a server less cloud platform to automatically build the API and serve the model to a web app. The web app has a simple user interface with an input field for the independent variables and an output field that displays the prediction result. For this research, only the final cost and concrete volume prediction models were deployed to show how the deployment can be done. The next flow chart describes the process.

The model built is first exported as a Predictive Model Markup Language (PMML) format which is then imported to the Clouderizer platform. Once the model is successfully imported, the input and output variables are identified and the API is automatically generated along with the web app. This generated web app can then be accessed using an automatically generated website address. The whole process is displayed in Figure 32.

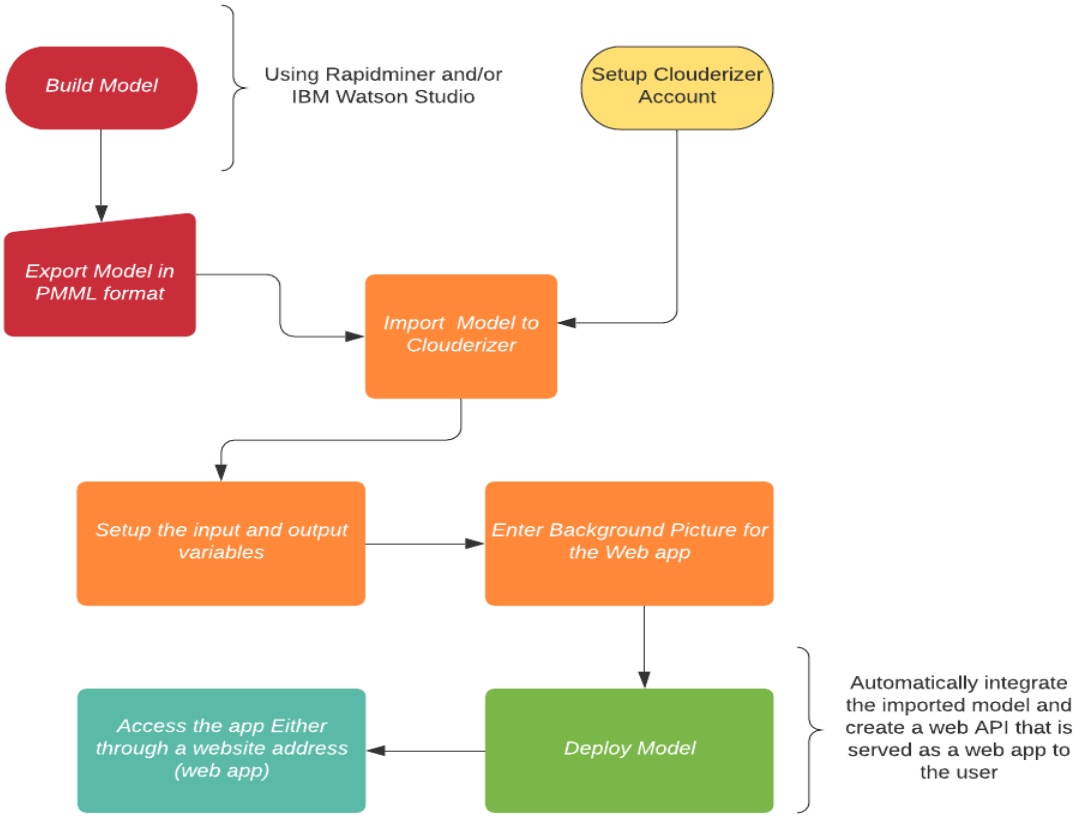

Figure 32: Process flow for Model Deployment

The web app can be accessed by entering the web address automatically generated for it. As the server only allows one user per session for a free account, the web app can only be accessed from the Clouderizer account. The web app has a simple user interface as shown in Figure 34. The values for the input variables are presented in the left side and the prediction result is presented on the right. The app gives the user two options when entering the input variables.

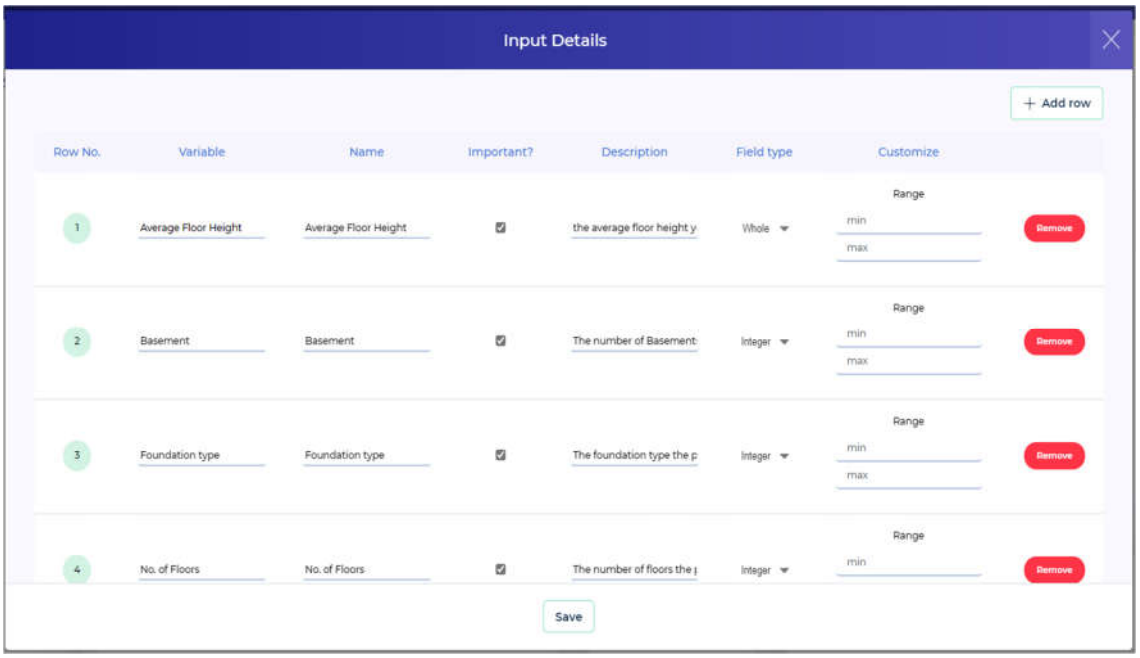

Figure 33: Identifying Input Parameters for the deployed model

The user can manually enter the values for each variable in the presented form or can import a csv file (located at the left bottom) if they wish to do multiple prediction at ones or when entering the values one at a time is tiresome, which could be the case when the independent variables are in tenth and sometimes hundreds.

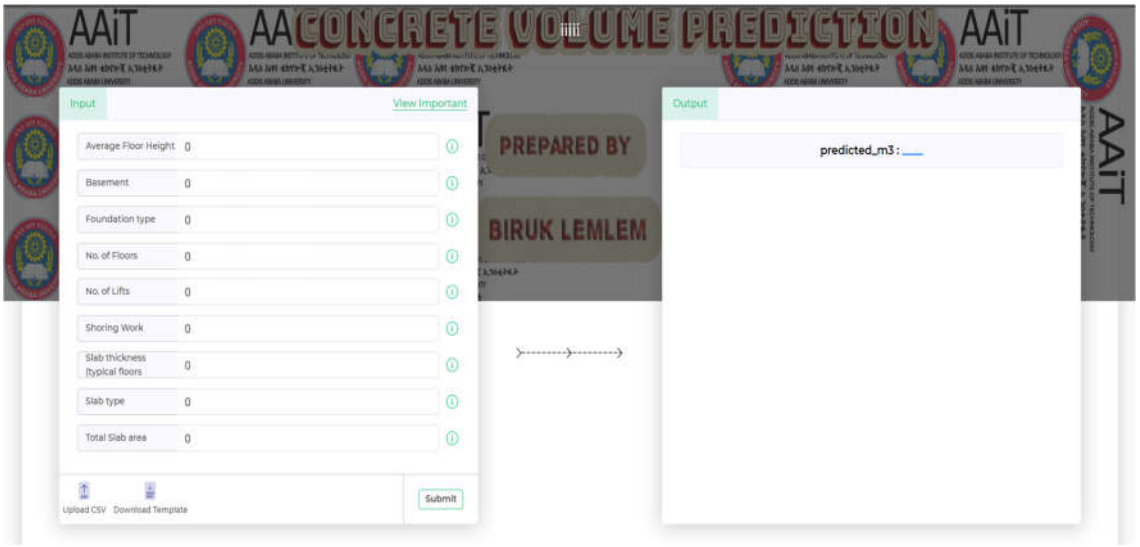

Figure 34: User interface of the web app

A short description on the attributes is presented for new users when they click the Information symbol next to each attribute. Once the values for the attributes are entered, one can make predictions by clicking on the submit button. The app will then call the imported model, perform the prediction and present the predicted value on the output box located on the right as shown in Figure 35.

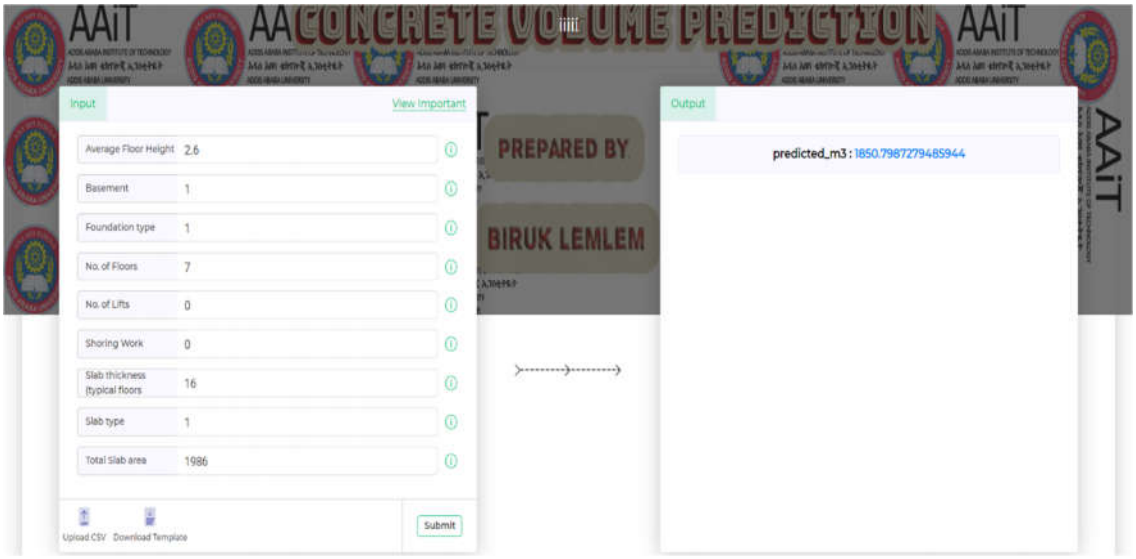

Figure 35: The App predicting and displaying the concrete volume for the given attribute values

6. CONCLUSION, RECOMMENDATION AND LIMITATION

6.1. CONCLUSION

Fast and reliable cost estimates of construction projects during the preliminary stages are becoming increasingly relevant in today's fast-paced, competitive environment. This need becomes even more apparent when considering the vast number of studies being conducted on this subject in various countries. Based on literature review, it was apparent there exists a gap in research with regards conceptual cost estimation of building projects in Ethiopia.

This study attempted to participate in filling this research gap by gathering data on buildings constructed in Addis Ababa to evaluate the accuracy of four powerful prediction algorithms in predicting the final cost of a building project. The Neural Network model was found to be superior of the four techniques with a relative error of 37.05%. This is in line with other researches that has also identified NNs to perform well in final cost estimation. While the relative error of the NN model was within the error margin for Class 4 estimate, it was found to be rather disappointing.

The accuracy of any data mining project ultimately depends on the quality of the data fed to the algorithm. The data fed to the NN algorithm in this case was the final cost of building projects executed over different time periods. As it's been pointed out in different literatures, adopting cost data in areas where construction costs are known to fluctuate dramatically would not provide a strong foundation for a prediction algorithm to learn on. Considering the price fluctuation seen in Ethiopia, this could be amongst the reasons for the lower accuracy that was gained from the Final cost prediction model. As a potential solution to this problem, an alternative to this another approach known as a quantity-based approach is proposed. The second objective of this study focused on determining if this approach is a better option to the cost-based one for the case of Structural cost prediction.

The quantity-based approach stems from the practices of quantity surveying where the cost of a work item is identified by multiplying the quantity of works and unit price for that item. Instead of using cost data, the quantity-based approach uses quantity of works data of building projects into prediction algorithms to predict the quantity of the respective work items. This effectively removes any cost fluctuations in the data from affecting the prediction model.

In order to compare between these approaches, a structural cost estimation model was first developed based on historical structural cost data. The performance of all four predictive models was again evaluated for this case. The gradient boosted trees were found to have superior performance for this case with a relative error of are 22.67%.

In order to determine the structural cost using the quantity - based approach, prediction models were built to determine the quantity for the three work items that make up the structural cost of a building. These are concrete work, Formwork and Reinforcement works. Once the quantity is determined, it can be multiplied with their respective unit rates and the products can be summed to find the structural cost.

Perhaps, what maybe new in this research has to do something about error propagation. When the cost for the individual work items is to be summed to get the structural cost, the errors in the individual work items will be propagated to the structural cost. Therefore, the propagated error for structural cost estimation was calculated for the quantity-based approach and is then compared with the error from the GBT model of the cost-based approach.

The result showed that even though the propagated error is larger than the errors seen for the individual work items, it is still far smaller than the one gained from predicting the structural cost using the cost-based approach. Thus, supporting the notion that the quantity-based approach can provide a more accurate cost estimate than the cost-based approach.

This along with the added benefit of being able to determine the material quantity required during the preliminary phase of the project makes the quantity-based approach superior to the cost based one. But it shall also be noted that this comes at a cost of developing different models for different work items – thereby making it more costly in terms of time, data and money.

With regards to the performance of the four algorithms, the following conclusions have been reached:

- For a small data set such as the ones collected for this research, linear regression and decision trees showed the least performance whereas Neural Networks and Gradient Boosted Trees showed good accuracies in all predictions.

- The Neural network and GBT have comparable performances indicating they can both be used for such prediction.
- While the importance of the independent variables differs based on what is to be predicted, the Total slab area, Number of floors, Basement and Average floor height were found to have the highest weight in most of the models. The total Slab area has the largest weight of them all indicating there is a strong relationship between the total slab area and the final and structural cost as well as the quantity of concrete, formwork and reinforcement.

6.2. LIMITATION

The main Limitation of this research is with regards to the availability of data. Some private institutions were not willing to provide data citing clients' privacy issues. The problem seen in government offices was poor handling of data. Most contracts and payment certificates are not backed up to a centralized database but rather left scattered in the hands of different parties. As a result of this, some of the soft copy backups seem to be gone with the engineers or managers when they leave office and accessing the hard copies of these documents was as difficult. As a result, it was not possible to gather more data within time.

In the beginning it was planned to assess the application of data mining for conceptual cost estimation of apartment buildings only, but as not enough apartment buildings data was gathered, other building types were included in order to increase the number of data. It would have been better for the quality of the research to focus on collecting data from a single institution or to choose randomly from a pool of contractors and consultants, but since not all are willing to provide their data, the researcher had to resort to collecting data on the basis of their willingness rather than random sampling.

6.3. RECOMMENDATION

Based on the research findings as well as knowledge gained throughout working on this research, the following recommendations are provided:

- Data mining algorithms like NN and GBTs are very much capable of predicting cost and quantity of construction work item, provided they are given large data set.
- Considering the unreasonable and unpredictable fluctuation in construction cost seen in the country, using past cost data does not seem to help in building accurate cost estimation models. in cases where this is an issue, the quantity-based approach will provide a more accurate estimate.
- The quantity-based method can be used by both private developers and government agencies participating in mass housing projects to determine the number of materials needed ahead of time.
- Hyper parameter optimization algorithms like Genetic algorithm are very handy to determine values for parameters but care shall be taken to make sure the algorithms are not stuck at local optima. A grid optimization is found to be robust to such problems but the time and resource constraint that comes with it shall be seriously considered.
- What seems to be the main bottleneck for these techniques be adopted in the construction industry, not just in Ethiopia but the world is the poor data management seen among parties involved in construction. Different researches have outlined that the construction industry is among the slowest industries to leverage the abundant data generated within it into good use. Thus, Measures shall be taken by the government, Consultant and Contractors to digitalize their data management system so that they can start leveraging their own data into knowledge that will aid them in business decisions.

6.4. FUTURE RESEARCHES

Future researches should focus on developing cost and quantity prediction models for building projects in Ethiopia by using a much larger data set, preferably more than 300. While there are a lot of research on neural networks in this area, there is only 2 published papers on Gradient boosted trees. Thus, more research is needed to corroborate the hopeful performance and potential use of Gradient boosted trees in conceptual cost estimations. Adopting GBTs will help us in building interpretable model as opposed to the "black-box" nature of NNs.

These data mining techniques using the quantity - based approach should also be adopted to road projects. The relative reduction in work items seen in road projects as compared to building projects will make it easier in developing a final cost estimation by using the quantity-based approach.

The main limitation to the quantity-based approach is we are required to develop different models for each work items. As this is can be a tiresome task, future research should also focus on developing multi – output prediction models that can be used to predict the quantity of works for different items at once.

REFERENCES

Abebe, C. (2017). *Determining Optimum Cost Contingency for Road Construction Project Using Monte Carlo Simulation: A Case of Road for Housing Projects.* Addis Ababa Inistitute of Technology, School of Civil and Environmental Engineering . Addis Ababa: Addis Ababa Univerisy.

Ahiaga-Dagbui, D. D. (2014). *Rethinking Construction Cost Overruns: An artificial Neural network appoach to construction cost estimation.* Edinburgh: The University of Edinburgh.

Ahiaga-Dagbui, D. D., & Smith, S. D. (2012). Neural Networks for modelling the final target cost of water projects. *Annual ARCOM confrence.* Edinburgh, United kingdom.

Alemayehu, S. (2014). *Testing Regression Models To Estimate Costs of Road Construction Projects.* Addis Ababa, Ethiopia: Addis Ababa Institute of Technology.

Alemu, A. (2020). *Hypothetical Modeling Of Contractor's Bid Markup Estimation For Road Construction Projects In Ethiopia.* Addis Ababa Institute Of Technology, School Of Civil And Environmental Engineering. Addis Ababa: Addis Ababa University.

Ali, R., & Rahinah, I. (2017). Industrialized Construction Chronology: The Disputes and Success Factors for a Resilient Construction Industry in Malaysia. *The Open Construction and Building Technology Journal*, 286 - 300.

American Association of Civil Engineers (AACE). (2005). *AACE international recomended practices.* AACE.

Amit Kumar, Y., Hasmat, M., & S.S., C. (2013). Selection of most relevant input parameters using WEKA for arti cial neural network based solar radiation prediction models. *Elsevier*, 510-519.

Arafa, M., & Alqedra, M. (2011). Early Stage Cost Estimation Of Building Construction Projects using Artificial Neural networks. *Journal of Artificial Intelligence*, 63-75.

Asteway, Y. (2008). *Study on the Effects of Unpredictable Price Fluctuation on the Capacity of Construction Contractors.* Addis Ababa: Addis Ababa University.

Ayele, B. (2019). *Assessment of Cost and Time Overrun in Addis Djibouti Railway Project.* Addis Ababa, Ethiopia: Addis Ababa University.

Bakhoum, M., Morcous, G., Taha, M., & El-Said, M. (1998). Estimation of Quantities and Cost of Prestressed Concrete over the Nile in Egypt. *Journal of Egyptian Society of Engineers*, 17-32.

Bayram, S., Ocal, M. E., Oral, E. L., & Atis, C. D. (2013). Comparison of multi layer perceptron (MLP) andradial basis function (RBF) for construction costestimation: the case of Turkey. *Journal of Civil Engineering and Management*, 1-11.

Bekele, B. (2017). *Factors Affecting Time And Cost Overruns In Housing Construction: The Case Of Addis Ketema Sub City Housing Development Project.* Addis Ababa, Ethiopia: St. Mary's University.

Bekele, D. (2017). *Factors Influencing Cost Overrun: The Case Of Water And Sanitation Construction Project In Addis Ababa In Partial Fulfillment Of The Requirements For The Award Of Master Of Arts Degree In Project Management.* Addis Ababa, Ethiopia: Addis Ababa University.

Belay, M. (2004). *Use Of Composite Concrte Slab System Using Hollow Blocks And Precast Slab/Beam Member.* Addis Ababa: Addis Ababa University.

Bhirud, A. N., & Ambrule, V. R. (2017). Design Stage Construction Cost Prediction Of Building Projects Using Artificial Neural Network. . *International Journal Of Engineering Sciences & Research Technology*, 314-325.

Borja, G. D., & Bryan, T. A. (2016). Preliminary Resource-based Estimates Combining Artificial intelligence Approaches and traditional Techniques . *Creative Construction Conference* (pp. 261-268). Zurich, Switzerland: ScienceDirect.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees.* California: Wadsworth.

Brighterion. (2017, January 18th). *A closer look at AI: Case-based reasoning*. Retrieved from Brighterion mastercard: Brighterioin.com/articial-intelligence-101-case-based-reasoning

Browniee, J. (2019, 12 7). *Weka Machine Learning*. Retrieved from Machine Learning Mastery : https://machinelearningmastery.com/category/weka-machine-learning/

Canadian Construction Association . (2012). *Guide to Cost Prediction in Construction: An Analysis of Issues Affecting the Accuracy of Construcion Cost Estimates.* Canada: Canadian construction Association.

Chakraborty, D., Hosam, E., Hazem, E., & Lilian, G. (2020). A novel Construction Cost prediction Model Using hybrid Natural and light Gradient Boosting. *Journal of Advanced Engineering Informatics, 46*, 1-10.

Chen, W., & Liew, J. R. (2002). *The Civil Engineering Handbook (New Directions in Civil Engineering 23).* (W. Chen, & J. R. Liew, Eds.) Boca Raton, Florida: CRC Press.

Cheng, M.-Y., Tsai, H.-C., & Hsieh, W.-S. (2008). Web-based Conceptual Cost Estimates for Construction Projects using Evolutionary Fuzzy Neural Interference Model. *Automation in Construction*, 164-172.

Cheng, M.-Y., Tsai, H.-C., & Sudjono, E. (2009). Evolutionary Fuzzy Hybrid Neural Network for Conceptual Cost Estimates in Construcion Projects. *26th International Symposium on Automation and Robotics in Construction* (pp. 512-520). Austin, Texas: ISARC.

Cho, H., G., K., G., K. J., Y., & Kim, G.-H. (2013). A Comparison Of Construction Cost Estimation Using Multiple Regression Analysis and Neural Network in Elementary School Projects. *Jornal of the Korea Institute of Building Construction*, 66-74.

Chou, J., Peng, M., Persad, K., & O'Connor, J. (2006). Quantity-based approach to preliminary cost estimates for highway projects. *Transportation Research Record: Journal of the Transportation Research Board, 1946*(1), 22-30.

Christopher, R. &. (2001). Expert Judgement in Cost Estimating: Modelling the Reasoning Process. *Concurrent Engineering, 9*(4), 271-284.

Comrey, A., & Lee, H. (1992). *A first Course in Factor analysis.* Hillsdale, NJ: Erlbaum.

Dalhoum, A. L., & Al-Rawi, M. (2019). Highe-Order Neural Networks are Equivalent to Ordinary Neural Networks. *Modern Applied Science, 13*(2), 228-240.

David, L., Margaret, E., & Anthony, H. (2006). Predicting Construction Cost Using Multiple Regression techniques. *ASCE: Journal of Construction Engineering and Management* , 750-758.

Dele Samuel, K. (2014). An assessment of current prelimnary cost estimating practice in Nigeria. *Journal of Environmental Design and Management* , 97-111.

Dominic D, A.-D. (2014). Dealing With Construction Cost Overruns Using Data Mining. *Taylor and Francis Journal, Construction Management Economics* .

Du, J., & Bormann, J. (2014). Improved Similaity Measure in Case-Based Reasoning with Global Sensitivity Analysis: An Example of Construction Quantity Estimating. *Journal of Computing in CIvil Engineering, 28*(6), 1-20.

Dursun, O., & Stoy, C. (2016). Conceptual Estimation of Construction Costs Using the Multistep Ahead Approach. *Journal of Construction Engineering and Management, 142*(9), 1-10. doi:10.1061/(asce)co.1943-7862.0001150

Elbetagi, E. (2015). *Cost Estimating Lecture Note.* Mensoura, Egypt: Mensoura University.

Elmousalami, H. H. (2019). Aritificial intelligence and paramettric construction cost Estimation modelinig: State-of-the Art Review. *Journal of Construction Engineering and management, 146*(1), 1- 30. doi:https://doi.org/10.1061/(ASCE)CO.1943-7862.0001678

Elmousalami, H. H. (2019). *Comparison of Artificial Intelligence Techniques for Project Conceptual Cost Prediction.* Egypt: Zagazig University.

ElProCus Technologies. (2020, October 12). *Fuzzy Logic - How does Fuzzy Logic work: Architecture and operation*. Retrieved from Fuzzy Logic – A Way to Achieve Control Based on Imprecise Inputs.

Emad, E., Ossama, H., Abdel-Razek, R., & El-Fitory, A. (2014). Conceptual Cost Estimate of Libyan Highway Projects Using Artificial Neural Network. *Journal of Engineering Research and Applications*, 56-66.

Emsley, M., Lowe, D., A.R., D., A, H., & A, H. (2002). Data Modelling and the Application of Neural Network approach to the prediction of total construction cost. *Construct, Manage and Econo*, 465-472.

Flyvbjerg, B. (2008). Curbing optimism bias and strategic misrepresentation in planning: reference class forecating in practice. *European Planning Studies, 16*(1), 3-21.

Flyvbjerg, B., & Skamris Holm, M. K. (2005). How (In)accurate Are Demand Forecasts in Public Works Projects?: The Case of Transportration. *Journal of the American Planning Association, 71*(2), 131-146.

Friedman, J. H. (1999). *Greedy Function Approximation: A gradient Boosting Machine.* IMS.

Gandhi, R. (2018, june 07). *Support Vector Machine — Introduction to Machine Learning Algorithms*. Retrieved october 13, 2020, from Medium: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

García de Soto, B. (2014). *A methodology to make accurate preliminary estimates of construction material quantities for construction projects.* Zurich: ETHZ Zurich.

Gardner, B. J. (2015). *Applying Artificial neural networks to top-down construction cost estimating of highway projects at the conceptual stage.* Iowa: Iowa State University.

Getaneh, G. T., & Sumati, V. (2020). Neuro-fuzzy systems in construction engineering and management research. *Automation in Construction*, 1-23.

Goran, P. N. (2011). Recent Research Work Resulting In IMS Building Technology Improvements. *CONSTRUCTII*, 21-24.

Guangli Feng, L. L. (2013). Application of Genetic Algorithm and Neural Network in Construction Cost Estimate. *Advanced Material Research, 756*(759), 3194-3198.

Gunaydın, M., & Dogan, Z. (2004). ural network approach for early cost estimation of structural systems of buildings. *International Journal of Project management*, 595-602.

Gupta, M. M., & Bukovsky, I. (2012). *Fundamentals of Higher Order Neural Networks for Modeling and Simulation.* beijing, china: the chinese academy of science.

Gwang-Hee, k., Shin, J.-M., Kim, S., & Yoonseok, s. (2013). Comparison of School Building Construction CostsEstimation Methods Using Regression Analysis, NeuralNetwork, and Support Vector Machine. *Journal of Building Construction and Planning Research*, 1-7.

Hailemariam, K., Eyob, M., Gudissa, D., & T.Quezon, E. (2020). Correlation Analysis of Factors Affecting Shoring Construction Techniques in Central Business District of Addis Ababa, Ethiopia. *Journal of Xidian University, 14*(10), 692-706.

Hamilton, H. (2018, July 9). *KDD Process/Primary Task of Data Mining*. Retrieved from Computer Science 831: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/2_tasks.html

Heaton, J. (2017, 06 01). *The number of Hidden Layers*. Retrieved from Heaton Research: https://www.heatonresearch.com/2017/06/01/hidden-layers.html

Hornik, K. (1991). Approximation Capabilities of Multilayer Feedforward Networks. *Neural Networks, 4*(2), 251-257. doi:10.1016/0893-6080(91)90009-T

hssay. (2018, August 20). *Data Science Stack Exchange*. Retrieved from High RMSE and MAE and Low MAPE: https://datascience.stackexchange.com/questions/37168/high-rmse-and-mae-and-low-mape

IDCCPCA. (2020). *Annual Report. 2014/15 - 2019/2020.* Addis Ababa: IDCCPCA.

Idowu, O. S. (2019). *Development of Conceptual Quantity Models for Building Envelopes and Structural Elements.* Hong Kong, Chinga: City University of Hong Kong.

Jason Brown, l. (2020, August 15). *Machine Learning Mastery*. Retrieved from: https://machinelearningmastery.com/configure-gradient-boosting-algorithm/

Ji, S.-H., Ahn, J., Lee, H.-S., & Han, K. (2019). Cost Estimation Model Using Modified Parameters for Construction Projects. *Advances in Civil Engineering*, 1-10.

Ji, S.-H., Park, M., & Lee, H.-S. (2011). Cost estimation model for building projects using case-based reasoning. *NRC Research Press*, 570-583.

Joseph, A. (2013). Quanity Based Active Schematic Estimating (Q-Base) Model. *KSCE Journal of Civil Engineering*, 9-21.

Josh, W. (2017). Decision Trees and Cost Estimating. *ICEAA Professional Development and Training workshop.* Sandiego, California: ICEAA.

Jung, S., Pyon, J.-H., Lee, H.-S., Park, M., Yooon, I., & Rho, J. (2020). Construction Cost Estimation Using a Case-Based Reasoning Hybrid Genetic Algorithm Based on Local Search Method. *Sustainablity*, 1-17.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision Under Risk. *Econometrica, 47*(2), 263-91.

Kaiser, H. (1974). An Index of Factorial Simplicity. *Psychometrika* , 31-36.

Kass, R., & Tinsley, H. (1979). Factory Analysis. *Jounal of Leisure Res.*, 120-138.

Kim, G.-H., Yoon, J.-E., An, S.-H., Cho, H.-H., & Kang, K.-I. (2004). Neural network model incorporating a genetic algorithm in estimating construction costs. *Building and Environment*, 1333-1340.

Kim, K., Kim, K., & Kang, C. (2009). Approimate Cost Estimating Model for PSC beam Bridge. *Korean Society of CIvil Engineers Journal of Civil Engineering, 13*(6), 377-388.

Kiran, R., & Issac, J. M. (2018). A Comparative Study of Solid Slab with Ribbed Slab in Bale Robe Town, Ethiopia. *International Journal of Modern Trends in Engineering and Research*, 87-94.

Koo, C.-W., Hong, T., Hyun, C.-T., Park, S., & Seo, J.-O. (2010). A study on the development of a cost model based on the owner's decision making at the early stages of a construction project. *International Journal of Strategic Property Management*, 121-137.

Law, J. (2010). *A dictionary of Accounting.* Oxford,United Kingdom: Oxford University press.

Lesniak, A., & Zima, K. (2018). Cost Calculation of Construction Projects Including Sustainablity Factors using the Case Based Reasoning Method. *MDPI*, 1-14.

Mahamid, I. (2016). Preliminary Estimate for Reinforcement Steel Quantity in Residential Buildings. *Organization, Technology and Management in Construction, 8*, 1405-1410.

Mitchell, M. (1998). *An Introduction to Genetic Algorithms.* Cambridge, Masachusetts : MIT press.

Mohammed, G. M. (2013). *Assessment of Price Escalation and Adjustment Problems on Federal Road Construction Projects.* Addis Ababa, Ethiopia: Addis Ababa University.

Montana, D., & Davis, L. (1989). Training feedforward neural networks using Genetic Algorithms. *Proceedings of the 11th International Joint conference on Artificial intelligence* (pp. 762-767). IJCAI.

N.Bhirud, A., & Vinayak, R. (2017). Pre-Design Stage Construction Cost Prediction of Building Projects Using Artificial Neural Network. *International Journal of Engineering Sciences and Research Technology*, 314-326.

Nichols, M. (2007). *Review of Highways Agency's Major Roads progremme: a report.* London, United Kingdom: Secretary of State of Transport.

Niu, S.-C. (2020, 10 09). *Introduction to Estimation.* Naveen Jindal School of Management. Dallas, Texas: The University of Texas at Dallas. Retrieved from Operation Research (OPRE-6301)

Nunnally, J. (1978). *Psychometric Theory.* New York: McGraw-Hill.

Nyoni, T. (2019, November). Cost Overrun Factors In Construction Industry: A Case Of Zimbabwe. *MPRA Paper*, 1-13.

Oh, C., Park, C., & Kim, K. (2013). An Approximate cost estimation model based on standard quantities of steel box girder bridge substructure. *Korean Society of Civil Engineers Journal of Civill Engineering, 17*(5), 877-885.

Oldfield, P. (2009). *An Historical Analysis of Energy Consumption in High rise Buildings.* The council on Tall buildngs and urban habitat.

Oluwafunmibi, S. I., & Lam, K. C. (2020). Conceptual Quantities Estimation Using Bootstrapped Support Vector regression Models. *Journal of Construction Engineering Management, 146*(4), 1-14. doi:10.1061/(ASCE)C).1943-7862.0001780

Pal, B., Mhashilkar, A., Pandey, A., Nagphase, B., & Chandanshive, V. (2015). Cost Estimation Model (CEM) of Buildings by ANN (Aritifical Neural Network). *International Advanced Research Journal in Science, Engineering and Technology, 5*(2), 26-28.

PASQUIRE, L. M. (1999). Examination of relationships between building form and function, and the cost of mechanical and electrical services. *Construction Management and Economics, 17*(4), 483-492.

Patil, M. P., & Salunkhe, M. A. (2020). Comparative Analysis Of Construction Cost Estimation Using Artificial Neural Networks. *Journal of Xidian University*, 1287-1305.

Peško, I., Mučenski, V., Šešlija, M., Radović, N., Vujkov, A., Bibić, D., & Milena Krklješ. (2017). Estimation of Costs and Durations of Construction of Urban Roads Using ANN and SVM. *Complexity*, 13.

Petruseva, S., Zileska-Pancovska, V., Žujo, V., & Brkan-Vejzović, A. (2017). Construction Costs Forecasting: Comparison of The Accuracy of Linear Regression and Support Vector Machine Models. *Technical Gazette, 24*(6), 1431-1438.

PMI. (2017). *A Guide to the Project Management Body of Knowledge (PMBOK GUIDE).* Newtown Square, Pennsylvania, United States: The Project Management Institute, PMI.

Roxas, C. L., & Ongpeng, J. M. (2014). An Artificial Neural Network Approach to Structural Cost Estimation of Building Projects in the Philippines. *DLSU Research Congress.* Manila, Philippines: De La Salle University.

Rudiger, W., & Jochen, H. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. *International Conference on the Practical Applications of Knowledge Discovery and Data Mining* .

Sawalhi, N. I. (2012). Modeling the Parametric Construction project Cost Estimate using Fuzzy Logic. *Intrernational Journal of Emerging Technology and Advanced Engineering, 2*(4), 1-8.

Senaviratna, N., & Cooray, T. (2019). Diagnosing Multicollinearity in logistic Regression Model. *Asian Journal of Probability and Statistics, 5*(2), 1-9.

Shehatto, O. M. (2013). *Cost Estimation for Building Construction Projects in Gaza Strip Using Artificial Neural Network (ANN).* Gaza: The Islamic University - Gaza Deanery of Graduate Studies .

Shin, Y. (2015). Application of Boosting Regression Trees to Preliminary Cost Estimation in Building Construction Projects. *Computational Intelligence and Neuroscience*. doi:https://doi.org/10.1155/2015/149702

Simone. (2012, 11 26). *StackExchange*. Retrieved from Mathematics behind classification and regression trees: https://stats.stackexchange.com/questions/44382/mathematics-behind-classification-and-regression-trees

Singh, S. (1990). Cost model for Reinforced concrete beam and Slab structures in Buildings. *Journal of Construction Engineering and Management*, 54-67.

Siqueira, I. (1998). Nural Network for Cost Estimating of Structural Steel Buildings . *AACE International Transactions*, 22-26.

Smita K., M., & Adamuthe, A. C. (2017). Construction Cost Prediction Using Neural Network. *ICTACT Journal of Soft Computing* , 1549-1556.

Sonmez, R. (2004). Conceptual Cost Estimation of Building Projects with regression Analysis and Neural Networks. *Canadiean Journal of Civil Engineering* , 677-683.

Sumathi, S., & Sivanandam, S. (2006, 12 10). *Introduction to Data Mining and its Applications.* Deblik, Berlin: Springer. Retrieved from Guru99: https://www.guru99.com/Data-Mining-Tutorial.html

T.Laroese, D., & D.Larose, C. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining.* River Street, Hoboken, NJ: John Wiley & Sons, Inc.

Tadesse, N., & Dinku, A. (2017). Conceptual Cost estimation of Road Projects In ethiopia. *Journal of EEA*, 17-29.

Tadesse, Y. (2018). *Assesment of Factors Affecting Cost Overrun in Public Building Construction Projects: Case Study at the Ethiopian Construction works Corporation (ECWC).* Addis Ababa, Ethiopia: Addis Ababa University.

Thomas, J., Coors, S., & Bernd, B. (2018). *Automatic Gradient Boosting.* Cornell University.

Trevor Hastie, R. T. (2017). *The Elements of Statistical Learning: Data mining, Inference and prediction.* Springer.

Vasily, K., & Edward, K. (1974, December). Predesign Cost-Estimation Function for Buildings. *Journal of the Construction Division, 100*(4), 589-604.

Vern, L. (2000). *Uncertainties and Error Propagation.* New York: Rochester Institute of Technology.

Wang, W.-C., Bilozerov, T., Dzeng, R.-J., Hsiao, F.-Y., & wang, K.-C. (2017). Conceptual Cost Estimations Using Neuro-Fuzzy And Multi-Factor Evaluation Methods For Building Projects. *Journal of Civil Engineering and Management, 23*(1), 1-14.

Wen-der, Y., & Skibniewski, M. J. (2010). Integrating Neurofuzzy System with Conceptual Cost Estimation to Discover Cost-Related Knowledge from Residential Construction Projects. *Journal of Computing in Civil Engineering, 24*(1).

Wendmu, N. (2018). *Causes Of Project Delay And Cost Overrun In Enyi Construction.* Addis Ababa, Ethiopia: Addis Ababa University.

Wllmott, C., & Matsuura, K. (2005). Advantages of the mean absolute error(MAE) over the Root mean squared error (RMSE). *Climate research, 30*(1), 79-82.

Wrike. (2020, 09 10). *What is cost Estimation in Project Management* . Retrieved from Wrike: Project Management Guide Overview: https://www.wrike.com/project-management-guide/faq/what-is-cost-estimation-in-project-management/

Wu, Y., Zhang, B., & Du, J. L.-L. (2011). Fuzzy Logic and Neuro-fuzzy Systems: A Systematic Introduction. *International Journal of Artificial Intelligence and Expert Systems (IJAE), 2*(2), 47-80.

Yadav, R., Vyas, M., Vyas, V., & Agrawal, S. (2016). Development of cost estimating method for bircklayer cost. *International Journal of Engineering Research and Technology*, 430-432.

Yeh, I. (1998). Qunaitty Estimating of Building using Logarithm-neuron Networks. *Journal of Construction Engineering and Management*, 374-380.

Zhang, Y. (2017). *Preliminary Cost Estimation Of Public Transportation Projects Using Data-Driven Based Method.* University of Florida.

Zima, K. (2015). The Case-Based Reasoning Model Of Cost Estimation At The Preliminary Stage of A Constructiono Project. *Procedia Engineering*, 57-64.

Zinabu, T. (2015). Causes of Contractor Cost Overrun in Construction Projects: The Case of Ethiopian Construction Sector. *International Journal of Business and Economics Research, 1*(4), 180-191.

APPENDIX

Table 34: sample data used for Final cost estimation

Building ID	Year	BT	ST	FT	IDF	ED	BST	SLT	NL	GNR	FS	DS	SW	Final Cost (ETB)	TSA (m^2)	BH (m)
GH17G1	2017	0	1	0	2	1	0	15	0	0	0	0	0	9,969,511.28	531	3.4
MH18G1	2018	1	1	0	1	1	0	16	0	0	0	0	0	24,169,303.00	1442.19	3.55
AR10G2	2010	3	1	0	2	1	0	15	0	0	0	0	0	3,704,249.12	361.8834	5.92
Pr19G2	2019	1	1	0	2	2	0	23	0	0	1	1	0	130,630,452.00	6298.53	7.6
KBA15G10	2015	2	0	0	2	1	1	30	2	1	1	0	0	63,405,681.06	4486.59	35.2
DCT15G11	2015	2	1	1	2	2	1	30	3	1	1	1	1	302,897,261.90	23,237.71	37.2
KM19G14	2019	2	1	1	1	1	1	20	2	0	0	0	1	233,046,492.50	10918.23	38.08
OT17G11	2019	2	0	1	2	2	3	30	2	1	1	0	1	247,462,375.20	6546.1	39.27

- *Shaded rows are detected as outliers and thus, removed*

Table 35: Gradient Boosted trees Cross Validation Result for Final Cost Estimation

Row no.	Actual Final Cost (ETB)	Predicted Final Cost (ETB)	Error	Absolute % error	Avg % error
1	21,970,996.54	21,736,933.54	234,063.00	1%	
2	66,827,845.72	53,314,731.67	13,513,114.05	20%	
3	3,704,249.12	6,389,538.71	(2,685,289.59)	72%	
4	46,208,213.11	58,440,187.30	(12,231,974.19)	26%	
5	53,982,597.41	57,591,791.01	(3,609,193.60)	7%	
6	47,053,627.75	58,157,545.06	(11,103,917.31)	24%	
7	33,273,001.50	18,981,237.38	14,291,764.12	43%	
8	54,080,695.68	34,893,021.47	19,187,674.21	35%	
9	63,405,681.06	178,482,807.14	(115,077,126.08)	181%	
10	26,108,087.61	42,728,590.89	(16,620,503.28)	64%	
11	75,241,098.35	62,175,036.29	13,066,062.06	17%	
12	4,560,960.64	6,545,420.79	(1,984,460.15)	44%	
13	41,747,614.00	41,935,491.84	(187,877.84)	0%	
14	54,085,671.88	65,511,699.90	(11,426,028.02)	21%	
15	5,943,386.65	4,487,992.82	1,455,393.82	24%	
16	83,938,178.67	75,123,985.97	8,814,192.70	11%	
17	23,728,070.94	17,462,512.73	6,265,558.21	26%	
18	4,755,283.84	9,365,319.16	(4,610,035.32)	97%	
19	247,462,375.20	110,565,466.81	136,896,908.39	55%	
20	5,991,490.29	10,404,609.89	(4,413,119.61)	74%	37%
21	65,709,116.57	33,267,653.15	32,441,463.42	49%	
22	52,820,082.40	48,268,999.86	4,551,082.54	9%	
23	233,046,492.50	120,456,811.54	112,589,680.96	48%	
24	6,618,201.17	4,480,896.42	2,137,304.75	32%	
25	30,182,992.27	46,970,225.47	(16,787,233.20)	56%	
26	14,252,279.78	23,107,575.77	(8,855,295.99)	62%	
27	24,169,303.00	9,678,606.36	14,490,696.64	60%	
28	9,355,712.08	4,761,582.01	4,594,130.07	49%	
29	56,449,022.83	52,434,527.98	4,014,494.85	7%	
30	90,497,239.27	56,041,714.82	34,455,524.45	38%	
31	24,948,142.09	16,875,186.13	8,072,955.96	32%	
32	21,415,713.91	18,972,084.50	2,443,629.41	11%	
33	45,844,339.71	44,817,363.06	1,026,976.65	2%	
34	19,472,609.37	17,426,090.16	2,046,519.21	11%	
35	54,895,218.07	45,816,552.99	9,078,665.08	17%	
36	9,969,511.28	16,254,930.18	(6,285,418.90)	63%	
37	23,873,740.52	20,475,181.35	3,398,559.17	14%	
38	75,145,651.31	83,902,194.00	(8,756,542.69)	12%	
39	44,196,169.59	49,975,954.62	(5,779,785.03)	13%	

Table 36: ANN Cross validation result for final cost estimation

Row no	Actual Final Cost (ETB)	Predicted Final Cost (ETB)	Error	Absolute % error	Average % error
1	47,053,627.75	36,310,660.50	10,742,967.25	23%	
2	52,820,082.40	28,914,354.01	23,905,728.39	45%	
3	130,630,452.00	122,050,712.06	8,579,739.94	7%	
4	24,948,142.09	16,082,508.63	8,865,633.46	36%	
5	30,182,992.27	24,783,610.38	5,399,381.89	18%	
6	14,252,279.78	18,054,107.24	(3,801,827.46)	27%	
7	65,709,116.57	21,466,111.75	44,243,004.82	67%	
8	5,991,490.29	4,222,178.81	1,769,311.48	30%	
9	56,449,022.83	51,951,708.07	4,497,314.76	8%	
10	24,169,303.00	8,737,264.64	15,432,038.36	64%	
11	66,827,845.72	35,510,122.71	31,317,723.01	47%	
12	46,208,213.11	63,312,372.54	(17,104,159.43)	37%	
13	23,728,070.94	27,760,356.00	(4,032,285.06)	17%	
14	26,108,087.61	41,797,351.45	(15,689,263.84)	60%	
15	41,747,614.00	68,071,491.54	(26,323,877.54)	63%	
16	23,873,740.52	15,128,125.66	8,745,614.86	37%	
17	4,755,283.84	4,453,072.30	302,211.54	6%	
18	90,497,239.27	56,411,782.92	34,085,456.35	38%	
19	83,938,178.67	79,866,343.11	4,071,835.56	5%	
20	4,560,960.64	9,555,032.03	(4,994,071.40)	109%	37%
21	63,405,681.06	65,800,198.44	(2,394,517.38)	4%	
22	33,273,001.50	15,250,468.95	18,022,532.55	54%	
23	75,241,098.35	84,747,225.25	(9,506,126.90)	13%	
24	75,145,651.31	91,717,242.94	(16,571,591.63)	22%	
25	21,415,713.91	24,240,820.02	(2,825,106.11)	13%	
26	21,970,996.54	40,151,330.44	(18,180,333.90)	83%	
27	9,969,511.28	9,876,136.32	93,374.96	1%	
28	19,472,609.37	37,654,683.53	(18,182,074.16)	93%	
29	9,355,712.08	1,398,279.51	7,957,432.57	85%	
30	54,085,671.88	63,033,978.22	(8,948,306.34)	17%	
31	6,618,201.17	1,923,842.28	4,694,358.89	71%	
32	54,895,218.07	60,214,431.02	(5,319,212.95)	10%	
33	53,982,597.41	28,713,299.10	25,269,298.31	47%	
34	5,943,386.65	6,066,725.55	(123,338.90)	2%	
35	3,704,249.12	2,493,389.91	1,210,859.21	33%	
36	233,046,492.50	336,220,610.90	(103,174,118.40)	44%	
37	45,844,339.71	14,678,871.56	31,165,468.15	68%	
38	247,462,375.20	185,725,652.58	61,736,722.62	25%	
39	54,080,695.68	44,765,896.63	9,314,799.05	17%	

Table 37: Regression Cross validation result for Final Cost estimation

Row no	Actual Final Cost (ETB)	Predicted Final Cost (ETB)	Error	Absolute % error	Average % error
1	21,970,996.54	61,753,516.68	(39,782,520.14)	181%	
2	66,827,845.72	98,047,734.93	(31,219,889.21)	47%	
3	3,704,249.12	2,330,717.40	1,373,531.72	37%	
4	46,208,213.11	51,696,006.72	(5,487,793.61)	12%	
5	53,982,597.41	79,789,735.94	(25,807,138.53)	48%	
6	47,053,627.75	72,436,151.38	(25,382,523.63)	54%	
7	33,273,001.50	45,207,533.66	(11,934,532.16)	36%	
8	54,080,695.68	41,078,829.13	13,001,866.55	24%	
9	63,405,681.06	135,877,096.49	(72,471,415.43)	114%	
10	26,108,087.61	42,190,157.94	(16,082,070.33)	62%	
11	75,241,098.35	61,398,860.07	13,842,238.28	18%	
12	4,560,960.64	(5,011,541.93)	9,572,502.57	210%	
13	41,747,614.00	29,062,937.20	12,684,676.80	30%	
14	54,085,671.88	137,767,036.31	(83,681,364.43)	155%	
15	5,943,386.65	6,836,473.29	(893,086.65)	15%	
16	83,938,178.67	53,671,915.70	30,266,262.97	36%	
17	23,728,070.94	18,287,334.29	5,440,736.65	23%	
18	4,755,283.84	16,720,646.24	(11,965,362.40)	252%	
19	247,462,375.20	101,766,807.99	145,695,567.21	59%	
20	5,991,490.29	5,650,472.36	341,017.92	6%	61%
21	65,709,116.57	65,076,399.04	632,717.53	1%	
22	52,820,082.40	5,426,927.03	47,393,155.37	90%	
23	233,046,492.50	182,000,553.46	51,045,939.04	22%	
24	6,618,201.17	9,380,382.01	(2,762,180.84)	42%	
25	30,182,992.27	31,439,933.90	(1,256,941.63)	4%	
26	14,252,279.78	12,889,242.70	1,363,037.08	10%	
27	24,169,303.00	(6,667,601.07)	30,836,904.07	128%	
28	9,355,712.08	9,498,927.72	(143,215.64)	2%	
29	56,449,022.83	68,423,210.38	(11,974,187.55)	21%	
30	90,497,239.27	96,165,802.04	(5,668,562.77)	6%	
31	24,948,142.09	87,451,732.12	(62,503,590.03)	251%	
32	21,415,713.91	28,482,830.28	(7,067,116.37)	33%	
33	45,844,339.71	41,121,009.55	4,723,330.16	10%	
34	19,472,609.37	26,683,801.22	(7,211,191.85)	37%	
35	54,895,218.07	74,693,614.82	(19,798,396.75)	36%	
36	9,969,511.28	(8,895,244.66)	18,864,755.94	189%	
37	23,873,740.52	35,774,435.91	(11,900,695.39)	50%	
38	75,145,651.31	50,103,369.79	25,042,281.52	33%	
39	44,196,169.59	49,712,825.91	(5,516,656.32)	12%	

Table 38: Decision Tree Cross Validation Result for Final Cost Estimation

Row no	Actual Final Cost (ETB)	Predicted Final Cost (ETB)	Error	Absolute % error	Average % error
1	21,970,996.54	21,538,798.07	432,198.47	2%	
2	66,827,845.72	54,085,671.88	12,742,173.84	19%	
3	3,704,249.12	5,943,386.65	(2,239,137.53)	60%	
4	46,208,213.11	78,108,309.44	(31,900,096.33)	69%	
5	53,982,597.41	45,020,254.65	8,962,342.76	17%	
6	47,053,627.75	55,267,347.36	(8,213,719.61)	17%	
7	33,273,001.50	14,252,279.78	19,020,721.72	57%	
8	54,080,695.68	28,145,539.94	25,935,155.74	48%	
9	63,405,681.06	233,046,492.50	(169,640,811.44)	268%	
10	26,108,087.61	30,182,992.27	(4,074,904.66)	16%	
11	75,241,098.35	66,827,845.72	8,413,252.63	11%	
12	4,560,960.64	6,184,359.37	(1,623,398.73)	36%	
13	41,747,614.00	65,709,116.57	(23,961,502.57)	57%	
14	54,085,671.88	66,827,845.72	(12,742,173.84)	24%	
15	5,943,386.65	4,560,960.64	1,382,426.01	23%	
16	83,938,178.67	75,193,374.83	8,744,803.84	10%	
17	23,728,070.94	19,298,064.05	4,430,006.89	19%	
18	4,755,283.84	9,355,712.08	(4,600,428.24)	97%	
19	247,462,375.20	233,046,492.50	14,415,882.70	6%	
20	5,991,490.29	9,355,712.08	(3,364,221.79)	56%	42%
21	65,709,116.57	33,273,001.50	32,436,115.07	49%	
22	52,820,082.40	45,844,339.71	6,975,742.69	13%	
23	233,046,492.50	90,497,239.27	142,549,253.23	61%	
24	6,618,201.17	4,736,198.80	1,882,002.37	28%	
25	30,182,992.27	53,944,648.39	(23,761,656.12)	79%	
26	14,252,279.78	22,371,593.80	(8,119,314.02)	57%	
27	24,169,303.00	9,969,511.28	14,199,791.72	59%	
28	9,355,712.08	4,658,122.24	4,697,589.84	50%	
29	56,449,022.83	47,053,627.75	9,395,395.08	17%	
30	90,497,239.27	47,053,627.75	43,443,611.52	48%	
31	24,948,142.09	33,273,001.50	(8,324,859.41)	33%	
32	21,415,713.91	21,970,996.54	(555,282.63)	3%	
33	45,844,339.71	44,196,169.59	1,648,170.12	4%	
34	19,472,609.37	23,728,070.94	(4,255,461.57)	22%	
35	54,895,218.07	53,450,389.04	1,444,829.03	3%	
36	9,969,511.28	24,169,303.00	(14,199,791.72)	142%	
37	23,873,740.52	33,273,001.50	(9,399,260.98)	39%	
38	75,145,651.31	87,217,708.97	(12,072,057.66)	16%	
39	44,196,169.59	45,844,339.71	(1,648,170.12)	4%	

Table 39: Decision Tree Cross validation result for Concrete Quantity estimation

Row no	Actual Concrete quantity (m^3)	Predicted Concrete quantity (m^3)	Error	Absolute % error	Average % error
1	612.576312	553.6785093	58.90	10%	
2	540.7689877	553.6785093	(12.91)	2%	
3	364.470515	553.6785093	(189.21)	52%	
4	510.6542	383.3174322	127.34	25%	
5	3189.99	2103.805177	1,086.18	34%	
6	1023.790656	770.5966667	253.19	25%	
7	2299.9084	2103.805177	196.10	9%	
8	1301.104349	2103.805177	(802.70)	62%	
9	241.9843614	292.3651767	(50.38)	21%	
10	230.7651943	292.3651767	(61.60)	27%	
11	142.53	292.3651767	(149.84)	105%	
12	1081.685795	1372.682135	(291.00)	27%	
13	248.8401917	234.2859111	14.55	6%	
14	1485.526	2674.120133	(1,188.59)	80%	
15	1257.1185	1161.810181	95.31	8%	
16	736.94	787.7051667	(50.77)	7%	
17	2216.947107	1326.985014	889.96	40%	
18	2139.0296	1326.985014	812.04	38%	
19	1795.561725	1326.985014	468.58	26%	
20	358.5563	523.1410626	(164.58)	46%	27%
21	1202	833.895164	368.10	31%	
22	1260.673456	1812.977176	(552.30)	44%	
23	406.228175	409.4239385	(3.20)	1%	
24	788.2655	1147.148232	(358.88)	46%	
25	911	729.6851667	181.31	20%	
26	518.3	383.3174322	134.98	26%	
27	1222.564051	1344.506484	(121.94)	10%	
28	1797	2279.906909	(482.91)	27%	
29	1897	2279.906909	(382.91)	20%	
30	1526.782	2279.906909	(753.12)	49%	
31	265.74	230.9059495	34.83	13%	
32	2532.462	2045.970182	486.49	19%	
33	624.735528	510.8231629	113.91	18%	
34	290.41	225.9719495	64.44	22%	
35	663.85	812.0685	(148.22)	22%	
36	518	574.0952069	(56.10)	11%	
37	1102.6404	1368.491214	(265.85)	24%	
38	394.346254	425.146225	(30.80)	8%	
39	374.139	425.146225	(51.01)	14%	

Table 40: Gradient Boosted Trees Cross Validation Result for concrete quantity estimation

Row no	Actual Concrete quantity (m^3)	Predicted Concrete quantity (m^3)	Error	Absolute % error	Average Absolute % error
1	612.576312	774.9481264	(162.37)	27%	
2	540.7689877	517.7834396	22.99	4%	
3	364.470515	453.0011576	(88.53)	24%	
4	510.6542	378.8833911	131.77	26%	
5	3189.99	2279.278395	910.71	29%	
6	1023.790656	826.2106015	197.58	19%	
7	2299.9084	2511.597972	(211.69)	9%	
8	1301.104349	1684.779918	(383.68)	29%	
9	241.9843614	266.8842148	(24.90)	10%	
10	230.7651943	286.3651076	(55.60)	24%	
11	142.53	286.9172006	(144.39)	101%	
12	1081.685795	1639.393678	(557.71)	52%	
13	248.8401917	206.8734959	41.97	17%	
14	1485.526	1889.347606	(403.82)	27%	
15	1257.1185	1105.452586	151.67	12%	
16	736.94	666.4104001	70.53	10%	
17	2216.947107	1644.249659	572.70	26%	
18	2139.0296	1526.783862	612.25	29%	
19	1795.561725	1050.462412	745.10	41%	
20	358.5563	429.0330666	(70.48)	20%	20.42%
21	1202	952.5927098	249.41	21%	
22	1260.673456	1280.266201	(19.59)	2%	
23	406.228175	295.9788125	110.25	27%	
24	788.2655	1015.295707	(227.03)	29%	
25	911	731.401329	179.60	20%	
26	518.3	467.3604739	50.94	10%	
27	1222.564051	1319.556496	(96.99)	8%	
28	1797	2114.156762	(317.16)	18%	
29	1897	2495.51842	(598.52)	32%	
30	1526.782	2139.019826	(612.24)	40%	
31	265.74	267.9978877	(2.26)	1%	
32	2532.462	2591.863235	(59.40)	2%	
33	624.735528	621.9189285	2.82	0%	
34	290.41	273.2809006	17.13	6%	
35	663.85	737.626698	(73.78)	11%	
36	518	570.5341071	(52.53)	10%	
37	1102.6404	1262.33375	(159.69)	14%	
38	394.346254	401.0874833	(6.74)	2%	
39	374.139	343.1708926	30.97	8%	

Table 41: Artificial Neural Network Cross validation result for Concrete Quantity estimation

Row no	Actual Concrete quantity (m^3)	Predicted Concrete quantity (m^3)	Error	Absolute % error	Average Absolute % error
1	612.576312	792.1495522	(179.57)	29%	16.41%
2	540.7689877	534.5626777	6.21	1%	
3	364.470515	406.7575761	(42.29)	12%	
4	510.6542	452.629575	58.02	11%	
5	3189.99	2217.314967	972.68	30%	
6	1023.790656	796.3278139	227.46	22%	
7	2299.9084	2373.536974	(73.63)	3%	
8	1301.104349	1741.094742	(439.99)	34%	
9	241.9843614	218.8577595	23.13	10%	
10	230.7651943	225.6283446	5.14	2%	
11	142.53	151.6399296	(9.11)	6%	
12	1081.685795	1594.041682	(512.36)	47%	
13	248.8401917	315.5299701	(66.69)	27%	
14	1485.526	1153.69698	331.83	22%	
15	1257.1185	1194.04541	63.07	5%	
16	736.94	673.3781503	63.56	9%	
17	2216.947107	1980.701224	236.25	11%	
18	2139.0296	1848.83378	290.20	14%	
19	1795.561725	1469.401062	326.16	18%	
20	358.5563	584.5949674	(226.04)	63%	
21	1202	1440.930929	(238.93)	20%	
22	1260.673456	1135.680156	124.99	10%	
23	406.228175	498.1231363	(91.89)	23%	
24	788.2655	919.6726929	(131.41)	17%	
25	911	578.3012139	332.70	37%	
26	518.3	446.2299355	72.07	14%	
27	1222.564051	1261.375798	(38.81)	3%	
28	1797	1899.68632	(102.69)	6%	
29	1897	1732.548924	164.45	9%	
30	1526.782	2030.384743	(503.60)	33%	
31	265.74	272.4814191	(6.74)	3%	
32	2532.462	2667.651967	(135.19)	5%	
33	624.735528	502.9972616	121.74	19%	
34	290.41	343.3048472	(52.89)	18%	
35	663.85	703.1927903	(39.34)	6%	
36	518	528.0376534	(10.04)	2%	
37	1102.6404	1284.512822	(181.87)	16%	
38	394.346254	408.4358088	(14.09)	4%	
39	374.139	446.9565723	(72.82)	19%	

Table 42: Linear Regression Cross Validation result for Concrete Quantity estimation

Row no	Actual Concrete quantity (m^3)	Predicted Concrete quantity (m^3)	Error	Absolute % error	Average Absolute % error
1	612	1538.929422	(926.93)	151%	32.44%
2	540	746.084213	(206.08)	38%	
3	364	574.9693029	(210.97)	58%	
4	510	316.1654785	193.83	38%	
5	966	734.929502	231.07	24%	
6	1023	930.7079562	92.29	9%	
7	2299	2108.489919	190.51	8%	
8	1301	1560.928376	(259.93)	20%	
9	241	211.8303903	29.17	12%	
10	230	187.5225901	42.48	18%	
11	142	12.47298183	129.53	91%	
12	1081	1698.451816	(617.45)	57%	
13	248	85.26949117	162.73	66%	
14	1485	1777.971752	(292.97)	20%	
15	1257	1454.367516	(197.37)	16%	
16	736	770.6641852	(34.66)	5%	
17	2216	1471.070825	744.93	34%	
18	2127	1109.598604	1,017.40	48%	
19	1795	1588.993515	206.01	11%	
20	358	455.2784893	(97.28)	27%	
21	1202	1235.337787	(33.34)	3%	
22	1260	1097.825526	162.17	13%	
23	406	464.7874335	(58.79)	14%	
24	788	1007.999433	(220.00)	28%	
25	911	226.4343353	684.57	75%	
26	518	136.1443447	381.86	74%	
27	1222	1440.250303	(218.25)	18%	
28	2139	1690.32403	448.68	21%	
29	1897	2173.732725	(276.73)	15%	
30	1526	1729.427009	(203.43)	13%	
31	265	444.3700633	(179.37)	68%	
32	1797	1693.62244	103.38	6%	
33	624	624.1649371	(0.16)	0%	
34	290	494.6885876	(204.69)	71%	
35	663	735.1698222	(72.17)	11%	
36	518	631.8415175	(113.84)	22%	
37	1102	1464.013272	(362.01)	33%	
38	394	418.6429135	(24.64)	6%	
39	374	284.956616	89.04	24%	

Table 43: Sample of Original data for Structural Cost and Concrete, Rebar and formwork quantity.

Building ID	Concrete QTY (m^3)	Formwork QTY (m^2)	Rebar QTY (KG)	**Structural Cost**
RC19G11	3,620.3	20,928.4	1,281,805.4	ETB 12,156,020.49
C19G7	2,540.5	12,035.9	317,889.1	ETB 7,131,631.68
18G4	578.1	4,421.5	69,182.4	ETB 1,558,400.19
PS11G4	358.6	2,513.1	47,072.4	ETB 1,852,601.00
S13G4	358.6	2,513.1	47,072.4	ETB 2,003,999.00
TV18G5	624.7	4,837.6	73,702.8	ETB 1,710,358.59
TV19G5	966.3	4,030.2	79,842.3	ETB 2,639,726.97
TV18G7	2,299.9	14,867.4	322,860.4	ETB 5,021,821.42
FT14G5	1,283.0	6,512.6	168,157.0	ETB 2,793,184.73
KBA15G10	1,904.3	12,087.1	272,997.4	ETB 7,993,870.51
KBB15G4	1,234.4	9,042.0	188,200.5	ETB 6,000,153.10
KBC15G4	1,023.8	5,879.9	135,085.6	ETB 3,552,001.19
LM16G6	1,523.0	8,867.6	265,953.6	ETB 6,518,503.47
MkU19G4	788.3	4,352.7	82,681.7	ETB 5,711,190.72
MF16G5	518.3	2,969.3	60,830.0	ETB 1,033,653.80
OT17G11	7,604.5	35,392.2	908,900.2	ETB 20,633,521.06
SGC19G21	55,630.5	205,267.8	4,410,090.3	ETB 240,523,391.51
MSE16G2	510.7	2,470.4	70,446.4	ETB 1,395,941.37
DOM17G2	290.4	1,574.7	30,696.0	ETB 663,117.33
DCT15G11	10,750.5	38,023.1	1,364,967.4	ETB 32,767,581.61

**shaded rows indicate outliers thus, not used in the model training*

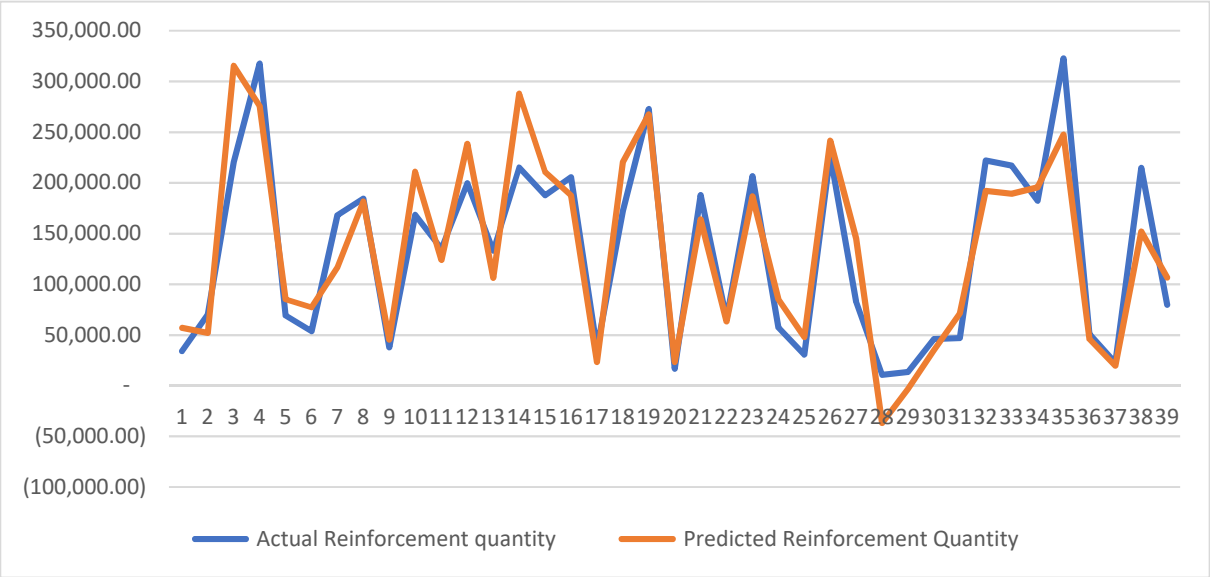

Figure 36: Actual Vs predicted Graph for Reinforcement work quantity prediction using LR

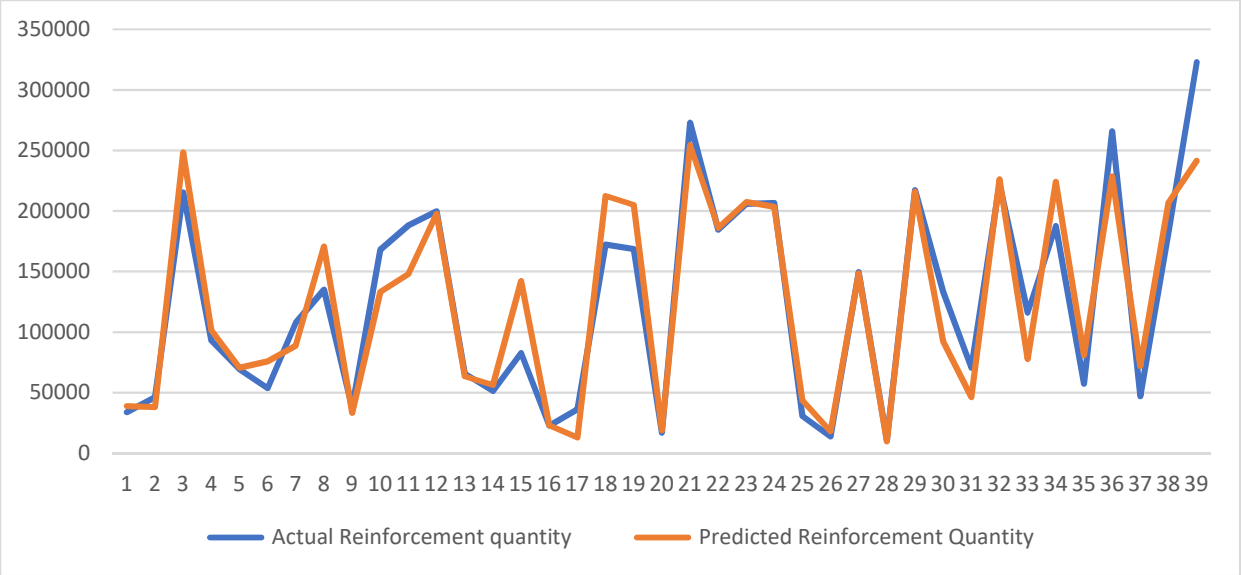

Figure 37: Actual Vs Predicted Graph for Reinforcement Work Quantity prediction using ANN

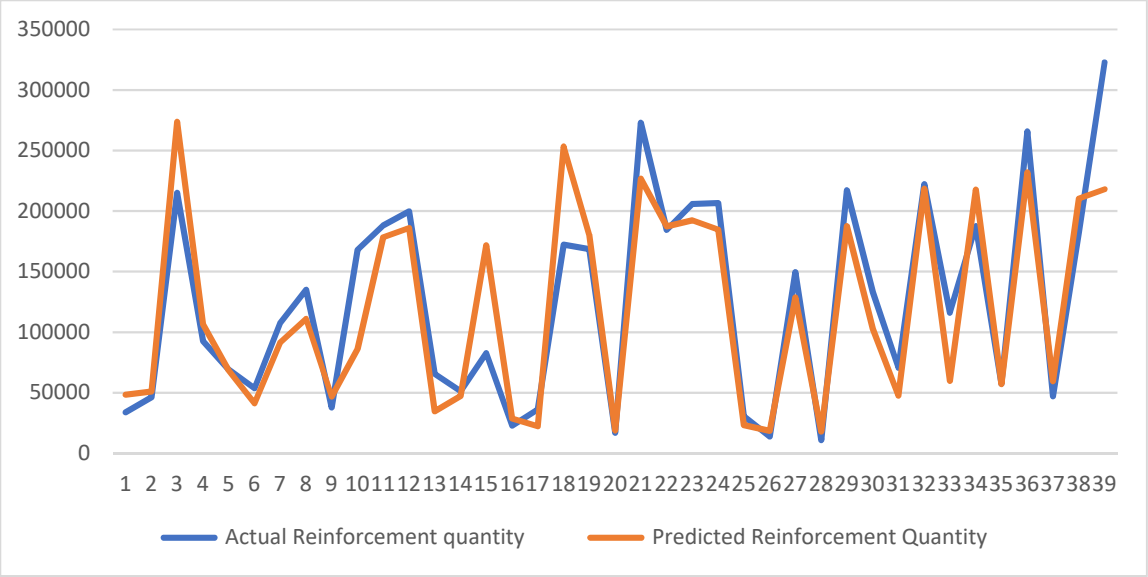

Figure 38: Actual Vs Predicted Graph for Reinforcement Quantity prediction using GBT

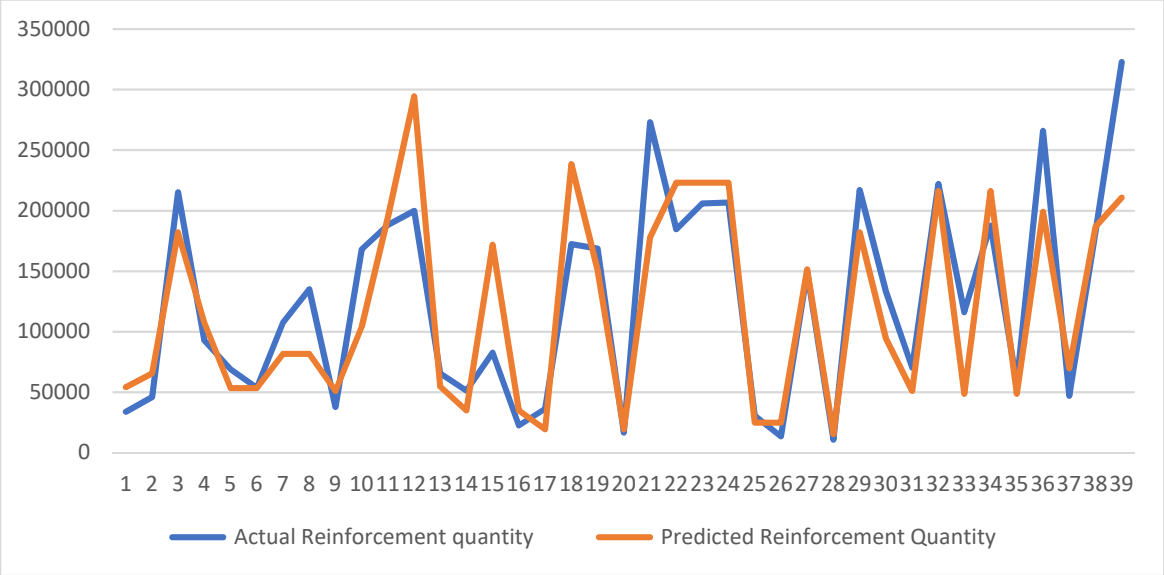

Figure 39: Actual Vs Predicted Graph for Reinforcement quantity Prediction using Decision Tree

Table 44: Decision Tree CV result for Formwork Quantity estimation

Row no	Actual Formwork quantity (m^2)	Predicted Formwork quantity (m^2)	Error	Absolute % error	Average Absolute % error
1	2629.0164	1469.930667	1,159.09	44%	
2	10043.99205	12061.47717	(2,017.49)	20%	
3	13384.40045	13766.252	(381.85)	3%	
4	7251.45	7771.5076	(520.06)	7%	
5	626.26	901.638	(275.38)	44%	
6	2533.95552	2629.0164	(95.06)	4%	
7	2894	2629.0164	264.98	9%	
8	3822.1916	2629.0164	1,193.18	31%	
9	3249.804	4148.880103	(899.08)	28%	
10	4000.9255	4148.880103	(147.95)	4%	
11	4698.7184	4148.880103	549.84	12%	
12	20928.42	17923.27	3,005.15	14%	
13	5879.92304	3890.481135	1,989.44	34%	
14	8260.113	9612.926633	(1,352.81)	16%	
15	8504	9612.926633	(1,108.93)	13%	
16	10357.3396	9612.926633	744.41	7%	
17	901.638	1212.411333	(310.77)	34%	
18	1398.818	1212.411333	186.41	13%	
19	4367.3175	3847.132374	520.19	12%	
20	17923.27	7598.155067	10,325.11	58%	22%
21	4421.522618	4063.342889	358.18	8%	
22	2800.165	7019.4244	(4,219.26)	151%	
23	12087.06382	10357.3396	1,729.72	14%	
24	13601.926	8867.6299	4,734.30	35%	
25	4259.07	3805.063462	454.01	11%	
26	4376.069	3805.063462	571.01	13%	
27	6806.313	6710.627747	95.69	1%	
28	1403.562	1503.115	(99.55)	7%	
29	4003.285	3880.226238	123.06	3%	
30	7257.168	6560.342747	696.83	10%	
31	12035.89052	10005.04735	2,030.84	17%	
32	1607.412	2685.657307	(1,078.25)	67%	
33	3792.705	3625.36475	167.34	4%	
34	7282.9022	7755.7815	(472.88)	6%	
35	8867.6299	10109.49653	(1,241.87)	14%	
36	6994.7922	6647.801347	346.99	5%	
37	10095.05	12035.89052	(1,940.84)	19%	
38	9876.1	12035.89052	(2,159.79)	22%	
39	13930.578	19425.845	(5,495.27)	39%	

Table 45: Gradient Boosted Tree CV result for Formwork Quantity estimation

Row no	Actual Formwork quantity (m^2)	Predicted Formwork quantity (m^2)	Error	Absolute % error	Average Absolute % error
1	2629.0164	2252.92456	376.09	14%	
2	10043.99205	12712.70714	(2,668.72)	27%	
3	13384.40045	10576.00397	2,808.40	21%	
4	7251.45	7805.147651	(553.70)	8%	
5	626.26	1099.141775	(472.88)	76%	
6	2533.95552	3403.656855	(869.70)	34%	
7	2894	3794.848386	(900.85)	31%	
8	3822.1916	4355.069561	(532.88)	14%	
9	3249.804	3092.562734	157.24	5%	
10	4000.9255	3592.391879	408.53	10%	
11	4698.7184	3842.504526	856.21	18%	
12	20928.42	16936.70371	3,991.72	19%	
13	5879.92304	3773.447018	2,106.48	36%	
14	8260.113	9039.854948	(779.74)	9%	
15	8504	9491.840869	(987.84)	12%	
16	10357.3396	9642.059708	715.28	7%	
17	901.638	1484.324871	(582.69)	65%	
18	1398.818	1761.117773	(362.30)	26%	
19	4367.3175	4242.163957	125.15	3%	
20	17923.27	14119.64933	3,803.62	21%	19.74%
21	4421.522618	4041.217016	380.31	9%	
22	2800.165	5247.65741	(2,447.49)	87%	
23	12087.06382	12727.68078	(640.62)	5%	
24	13601.926	11726.95914	1,874.97	14%	
25	4259.07	4778.957724	(519.89)	12%	
26	4376.069	4919.982196	(543.91)	12%	
27	6806.313	7895.044315	(1,088.73)	16%	
28	1403.562	1564.560209	(161.00)	11%	
29	4003.285	4020.219425	(16.93)	0%	
30	7257.168	6914.573629	342.59	5%	
31	12035.89052	8157.941238	3,877.95	32%	
32	1607.412	2080.240381	(472.83)	29%	
33	3792.705	3508.547543	284.16	7%	
34	7282.9022	7682.791919	(399.89)	5%	
35	8867.6299	11977.25856	(3,109.63)	35%	
36	6994.7922	7177.982537	(183.19)	3%	
37	10095.05	10635.59826	(540.55)	5%	
38	9876.1	10595.32654	(719.23)	7%	
39	13930.578	16348.94132	(2,418.36)	17%	

Table 46: ANN CV result for Formwork Quantity Estimation

Row no	Actual Formwork quantity (m^2)	Predicted Formwork quantity (m^2)	Error	Absolute % error	Average Absolute % error
1	2629.0164	2029.69971	599.32	23%	
2	10043.99205	12159.23687	(2,115.24)	21%	
3	13384.40045	11665.30086	1,719.10	13%	
4	7251.45	9317.782289	(2,066.33)	28%	
5	626.26	693.1799844	(66.92)	11%	
6	2533.95552	2326.525156	207.43	8%	
7	2894	2887.592384	6.41	0%	
8	3822.1916	3479.845954	342.35	9%	
9	3249.804	2505.565357	744.24	23%	
10	4000.9255	3308.525709	692.40	17%	
11	4698.7184	3588.388324	1,110.33	24%	
12	20928.42	17158.93275	3,769.49	18%	
13	5879.92304	5943.388154	(63.47)	1%	
14	8260.113	8018.809291	241.30	3%	
15	8504	9565.903499	(1,061.90)	12%	
16	10357.3396	11380.89186	(1,023.55)	10%	
17	901.638	1260.109056	(358.47)	40%	
18	1398.818	1249.144079	149.67	11%	
19	4367.3175	4719.68784	(352.37)	8%	
20	17923.27	15466.0034	2,457.27	14%	23.50%
21	4421.522618	4170.290239	251.23	6%	
22	2800.165	4798.385368	(1,998.22)	71%	
23	12087.06382	11048.75463	1,038.31	9%	
24	13601.926	14092.26302	(490.34)	4%	
25	4259.07	11849.98244	(7,590.91)	178%	
26	4376.069	4944.249949	(568.18)	13%	
27	6806.313	6495.468814	310.84	5%	
28	1403.562	2876.474972	(1,472.91)	105%	
29	4003.285	4368.16342	(364.88)	9%	
30	7257.168	5546.490328	1,710.68	24%	
31	12035.89052	8379.472763	3,656.42	30%	
32	1607.412	2179.644543	(572.23)	36%	
33	3792.705	3019.560108	773.14	20%	
34	7282.9022	9566.018532	(2,283.12)	31%	
35	8867.6299	10837.37609	(1,969.75)	22%	
36	6994.7922	5767.824557	1,226.97	18%	
37	10095.05	9961.135418	133.91	1%	
38	9876.1	10253.77412	(377.67)	4%	
39	13930.578	19139.42721	(5,208.85)	37%	

Table 47: Linear Regression CV result for Formwork Quantity Estimation

Row no	Actual Formwork quantity (m^2)	Predicted Formwork quantity (m^2)	Error	Absolute % error	Average Absolute % error
1	2629.0164	2415.501264	213.52	8%	
2	10043.99205	10484.40809	(440.42)	4%	
3	13384.40045	10312.95921	3,071.44	23%	
4	7251.45	10213.66204	(2,962.21)	41%	
5	626.26	982.4852244	(356.23)	57%	
6	2533.95552	2882.563983	(348.61)	14%	
7	2894	3336.564026	(442.56)	15%	
8	3822.1916	3080.856209	741.34	19%	
9	3249.804	3761.524442	(511.72)	16%	
10	4000.9255	2455.639276	1,545.29	39%	
11	4698.7184	2881.440306	1,817.28	39%	
12	20928.42	18198.31801	2,730.10	13%	
13	5879.92304	5564.988055	314.93	5%	
14	8260.113	8140.71699	119.40	1%	
15	8504	8507.535143	(3.54)	0%	
16	10357.3396	12126.9242	(1,769.58)	17%	
17	901.638	1270.347859	(368.71)	41%	
18	1398.818	1509.530261	(110.71)	8%	
19	4367.3175	3367.806519	999.51	23%	
20	17923.27	15363.87436	2,559.40	14%	25.48%
21	4421.522618	4058.607406	362.92	8%	
22	2800.165	4291.261293	(1,491.10)	53%	
23	12087.06382	11878.73794	208.33	2%	
24	13601.926	12151.83165	1,450.09	11%	
25	4259.07	8389.033485	(4,129.96)	97%	
26	4376.069	5701.131826	(1,325.06)	30%	
27	6806.313	6985.791779	(179.48)	3%	
28	1403.562	3547.454005	(2,143.89)	153%	
29	4003.285	3101.022305	902.26	23%	
30	7257.168	6002.413465	1,254.75	17%	
31	12035.89052	8999.792231	3,036.10	25%	
32	1607.412	1883.968681	(276.56)	17%	
33	3792.705	3251.132382	541.57	14%	
34	7282.9022	10751.3695	(3,468.47)	48%	
35	8867.6299	10525.47045	(1,657.84)	19%	
36	6994.7922	5928.881252	1,065.91	15%	
37	10095.05	10892.03564	(796.99)	8%	
38	9876.1	11172.10858	(1,296.01)	13%	
39	13930.578	19556.81057	(5,626.23)	40%	

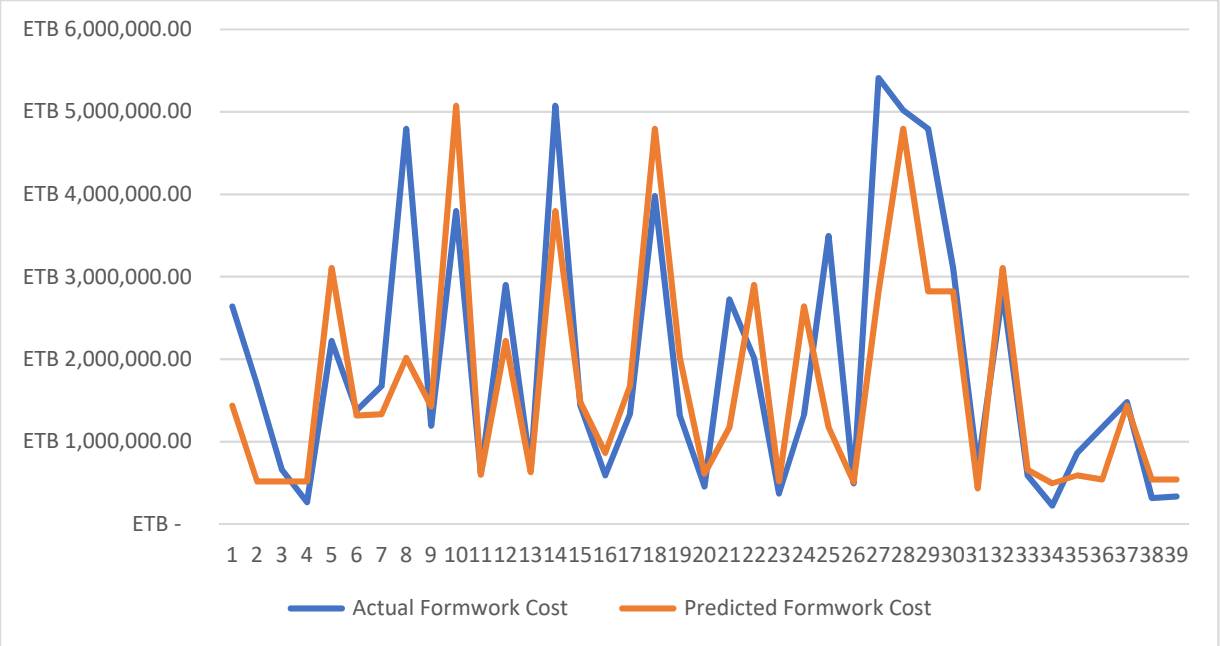

Figure 40: Actual vs Predicted Graph for Formwork Cost Using Decision Tree

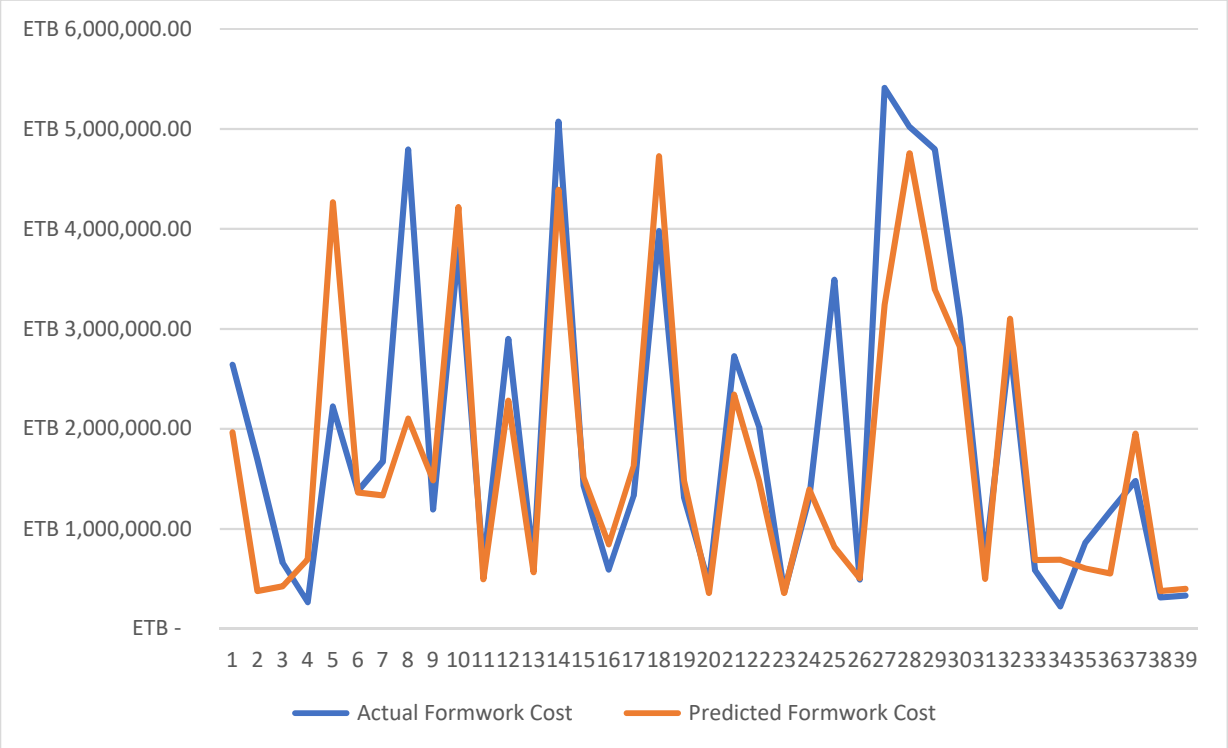

Figure 41: Actual Vs Predicted Graph for Formwork Cost using Gradient Boosted Trees

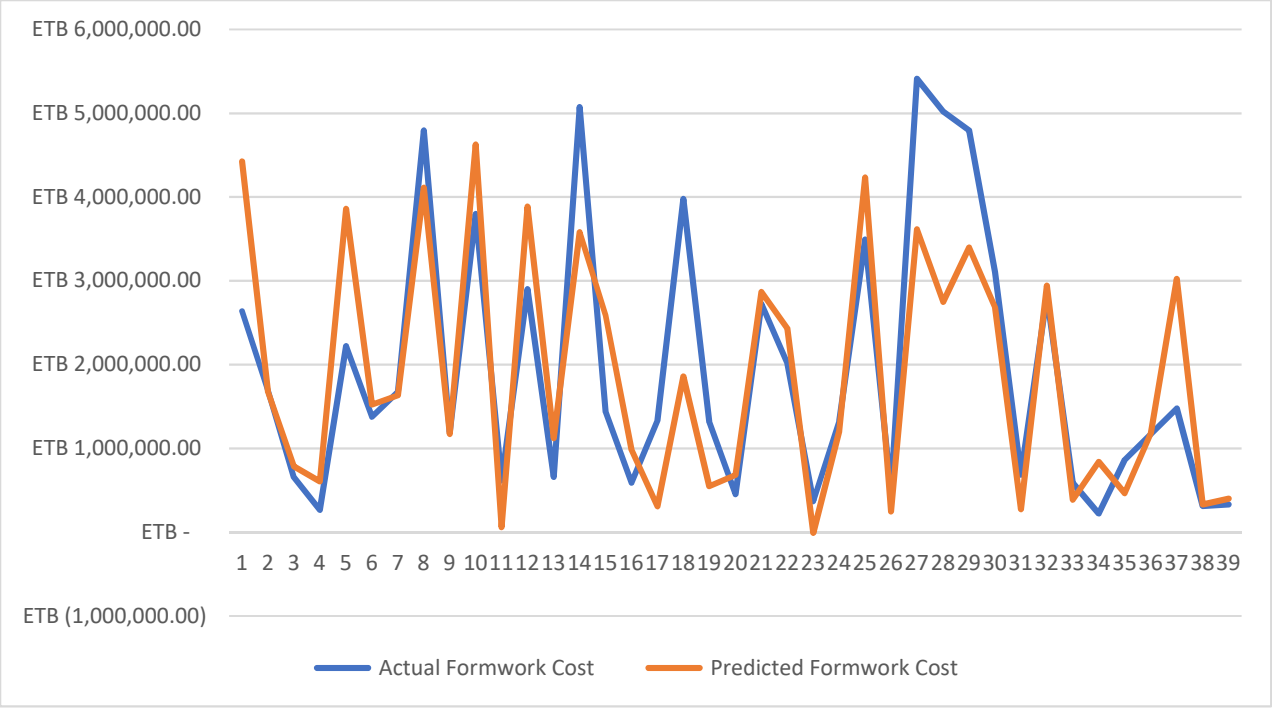

Figure 42: Actual Vs Predicted Graph for Formwork Cost using ANN

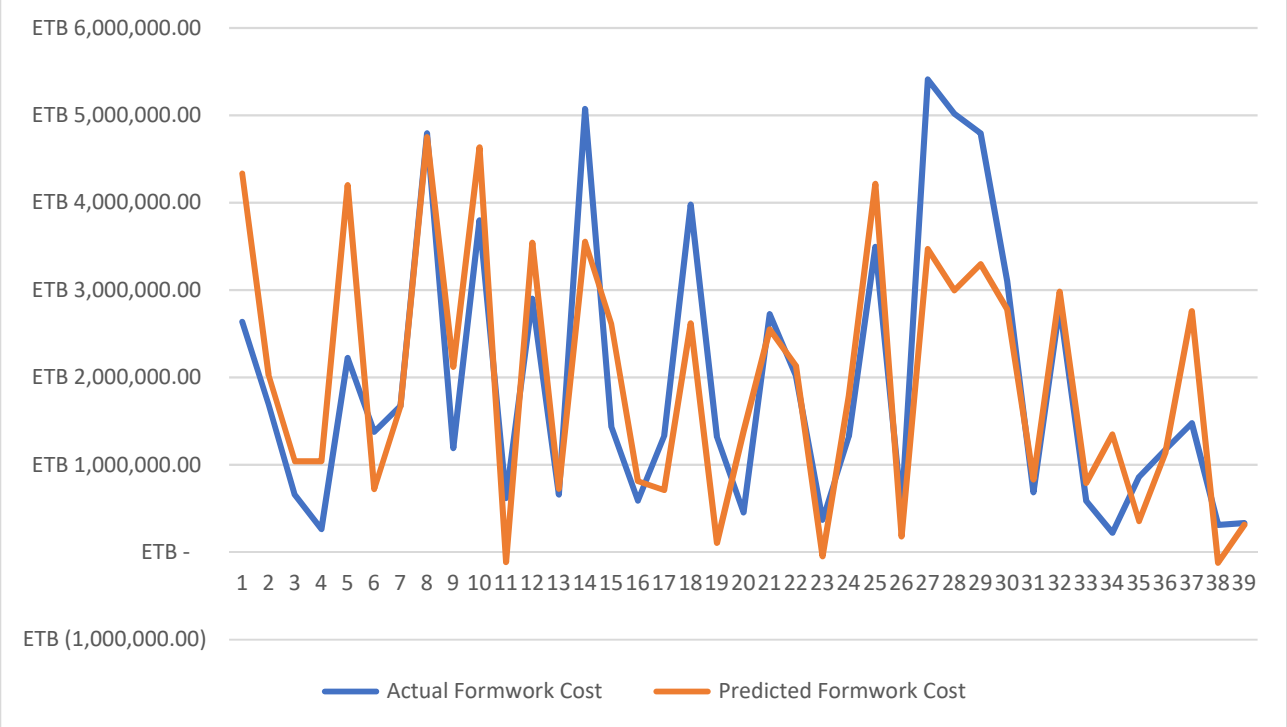

Figure 43: Actual Vs Predicted Graph for Formwork cost using Linear Regression

Table 48: Decision tree CV result for Structural Cost Estimation

Row no	Actual Structural Cost (ETB)	Predicted Structural Cost (ETB)	Error	Absolute % error	Average Absolute % error
1	3,726,700.72	3,958,306.78	(231,606.06)	6%	
2	24,256,775.58	29,562,971.71	(5,306,196.13)	22%	
3	6,594,113.30	11,819,618.09	(5,225,504.79)	79%	
4	4,859,836.85	3,958,306.78	901,530.07	19%	
5	7,707,032.77	2,902,268.16	4,804,764.61	62%	
6	2,676,461.89	4,056,386.67	(1,379,924.78)	52%	
7	12,478,276.57	12,781,638.84	(303,362.27)	2%	
8	3,955,407.78	3,488,188.82	467,218.96	12%	
9	8,119,134.94	9,149,702.55	(1,030,567.61)	13%	
10	3,804,284.70	3,670,505.43	133,779.27	4%	
11	16,195,294.97	6,235,416.94	9,959,878.03	61%	
12	17,806,452.37	11,883,136.67	5,923,315.70	33%	
13	12,220,832.23	9,944,045.12	2,276,787.11	19%	
14	11,387,584.90	16,583,899.44	(5,196,314.54)	46%	
15	4,973,374.71	3,872,020.80	1,101,353.92	22%	
16	15,613,615.70	20,655,232.97	(5,041,617.27)	32%	
17	3,791,727.89	2,793,880.82	997,847.07	26%	
18	35,451,341.78	17,288,957.62	18,162,384.17	51%	
19	36,870,783.50	16,346,714.44	20,524,069.06	56%	
20	2,893,236.56	2,793,880.82	99,355.74	3%	27%
21	22,589,408.08	18,955,046.67	3,634,361.41	16%	
22	23,986,009.78	24,522,831.95	(536,822.17)	2%	
23	17,079,813.18	18,955,046.67	(1,875,233.49)	11%	
24	8,126,358.30	8,123,584.86	2,773.44	0%	
25	6,489,290.50	2,826,999.40	3,662,291.10	56%	
26	29,248,493.16	27,067,112.92	2,181,380.25	7%	
27	6,488,621.01	4,056,386.67	2,432,234.34	37%	
28	2,911,299.76	3,968,167.07	(1,056,867.32)	36%	
29	16,771,462.86	11,932,930.74	4,838,532.13	29%	
30	7,981,174.18	8,171,979.57	(190,805.39)	2%	
31	11,783,548.53	11,932,930.74	(149,382.21)	1%	
32	2,946,137.84	3,841,054.25	(894,916.40)	30%	
33	7,042,727.07	11,670,080.17	(4,627,353.10)	66%	
34	15,173,783.08	11,883,136.67	3,290,646.41	22%	
35	24,788,888.32	29,562,971.71	(4,774,083.39)	19%	
36	22,296,477.64	6,818,420.19	15,478,057.45	69%	
37	4,308,488.64	3,502,437.45	806,051.19	19%	
38	29,877,450.25	26,752,634.37	3,124,815.88	10%	
39	8,270,445.47	8,075,555.80	194,889.67	2%	

Table 49: Gradient boosted trees CV results for Structural Cost Prediction

Row no	Actual Structural Cost (ETB)	Predicted Structural Cost (ETB)	Error	Absolute % error	Average Absolute % error
1	3,726,700.72	3,841,931.91	(115,231.20)	3%	
2	24,256,775.58	30,014,336.58	(5,757,561.00)	24%	
3	6,594,113.30	10,640,290.70	(4,046,177.40)	61%	
4	4,859,836.85	4,597,352.87	262,483.98	5%	
5	7,707,032.77	3,504,403.38	4,202,629.39	55%	
6	2,676,461.89	4,222,315.23	(1,545,853.34)	58%	
7	12,478,276.57	12,482,544.18	(4,267.61)	0%	
8	3,955,407.78	2,977,570.02	977,837.76	25%	
9	8,119,134.94	7,968,758.06	150,376.87	2%	
10	3,804,284.70	4,308,528.42	(504,243.73)	13%	
11	16,195,294.97	7,791,780.87	8,403,514.10	52%	
12	17,806,452.37	12,768,720.94	5,037,731.43	28%	
13	12,220,832.23	8,273,558.17	3,947,274.06	32%	
14	11,387,584.90	15,001,917.36	(3,614,332.46)	32%	
15	4,973,374.71	5,429,466.22	(456,091.51)	9%	
16	15,613,615.70	17,079,782.34	(1,466,166.64)	9%	
17	3,791,727.89	2,707,518.04	1,084,209.85	29%	
18	35,451,341.78	18,115,851.25	17,335,490.53	49%	
19	36,870,783.50	17,106,131.27	19,764,652.23	54%	
20	2,893,236.56	2,751,192.52	142,044.04	5%	23%
21	22,589,408.08	19,464,009.95	3,125,398.13	14%	
22	23,986,009.78	23,853,428.93	132,580.85	1%	
23	17,079,813.18	15,613,600.59	1,466,212.59	9%	
24	8,126,358.30	7,215,791.76	910,566.54	11%	
25	6,489,290.50	3,000,429.30	3,488,861.21	54%	
26	29,248,493.16	30,596,995.90	(1,348,502.74)	5%	
27	6,488,621.01	5,139,252.76	1,349,368.24	21%	
28	2,911,299.76	3,918,140.16	(1,006,840.40)	35%	
29	16,771,462.86	13,423,227.78	3,348,235.08	20%	
30	7,981,174.18	6,727,857.82	1,253,316.35	16%	
31	11,783,548.53	12,471,092.43	(687,543.90)	6%	
32	2,946,137.84	3,774,192.72	(828,054.87)	28%	
33	7,042,727.07	6,594,109.47	448,617.60	6%	
34	15,173,783.08	11,908,302.63	3,265,480.45	22%	
35	24,788,888.32	28,569,065.68	(3,780,177.36)	15%	
36	22,296,477.64	14,193,795.50	8,102,682.14	36%	
37	4,308,488.64	3,520,033.74	788,454.90	18%	
38	29,877,450.25	24,598,425.39	5,279,024.86	18%	
39	8,270,445.47	8,518,819.06	(248,373.59)	3%	

Table 50: ANN CV result for Structural Cost Estimation

Row no	Actual Structural Cost (ETB)	Predicted Structural Cost (ETB)	Error	Absolute % error	Average Absolute % error
1	3,726,700.72	981,568.12	2,745,132.60	74%	
2	24,256,775.58	27,615,204.17	(3,358,428.59)	14%	
3	6,594,113.30	6,461,112.81	133,000.50	2%	
4	4,859,836.85	4,023,956.01	835,880.84	17%	
5	7,707,032.77	12,521,688.57	(4,814,655.80)	62%	
6	2,676,461.89	6,503,532.25	(3,827,070.37)	143%	
7	12,478,276.57	17,156,212.10	(4,677,935.53)	37%	
8	3,955,407.78	1,278,987.71	2,676,420.07	68%	
9	8,119,134.94	16,101,022.86	(7,981,887.92)	98%	
10	3,804,284.70	4,269,222.07	(464,937.38)	12%	
11	16,195,294.97	20,210,620.57	(4,015,325.60)	25%	
12	17,806,452.37	31,396,724.96	(13,590,272.59)	76%	
13	12,220,832.23	6,287,434.28	5,933,397.95	49%	
14	11,387,584.90	10,338,252.58	1,049,332.32	9%	
15	4,973,374.71	5,253,325.38	(279,950.67)	6%	
16	15,613,615.70	24,021,575.05	(8,407,959.35)	54%	
17	3,791,727.89	5,840,958.73	(2,049,230.84)	54%	
18	35,451,341.78	23,212,919.84	12,238,421.94	35%	
19	36,870,783.50	23,968,372.68	12,902,410.82	35%	
20	2,893,236.56	5,136,239.31	(2,243,002.75)	78%	35.0%
21	22,589,408.08	22,501,749.00	87,659.08	0%	
22	23,986,009.78	22,159,392.75	1,826,617.03	8%	
23	17,079,813.18	22,644,978.48	(5,565,165.30)	33%	
24	8,126,358.30	11,186,470.03	(3,060,111.73)	38%	
25	6,489,290.50	7,496,069.07	(1,006,778.56)	16%	
26	29,248,493.16	25,453,995.10	3,794,498.06	13%	
27	6,488,621.01	8,846,337.57	(2,357,716.57)	36%	
28	2,911,299.76	2,253,359.53	657,940.23	23%	
29	16,771,462.86	17,662,753.67	(891,290.81)	5%	
30	7,981,174.18	5,127,879.96	2,853,294.21	36%	
31	11,783,548.53	18,556,883.25	(6,773,334.72)	57%	
32	2,946,137.84	2,886,898.58	59,239.27	2%	
33	7,042,727.07	6,780,381.22	262,345.85	4%	
34	15,173,783.08	24,349,392.13	(9,175,609.05)	60%	
35	24,788,888.32	24,681,645.15	107,243.17	0%	
36	22,296,477.64	17,065,388.07	5,231,089.57	23%	
37	4,308,488.64	5,798,827.11	(1,490,338.47)	35%	
38	29,877,450.25	23,648,336.53	6,229,113.72	21%	
39	8,270,445.47	7,460,400.21	810,045.26	10%	

Table 51: Linear Regression CV result for Structural Cost Estimation

Row no	Actual Structural Cost (ETB)	Predicted Structural Cost (ETB)	Error	Absolute % error	Average Absolute % error
1	2,946,137.84	4,107,653.08	(1,161,515.24)	39%	
2	24,256,775.58	28,199,843.43	(3,943,067.85)	16%	
3	6,594,113.30	7,456,506.48	(862,393.17)	13%	
4	4,859,836.85	7,760,556.55	(2,900,719.70)	60%	
5	7,707,032.77	14,636,470.48	(6,929,437.72)	90%	
6	3,726,700.72	136,499.75	3,590,200.96	96%	
7	7,042,727.07	4,311,383.96	2,731,343.11	39%	
8	3,955,407.78	(824,219.07)	4,779,626.85	121%	
9	6,489,290.50	6,462,976.29	26,314.22	0%	
10	16,771,462.86	14,009,694.29	2,761,768.57	16%	
11	16,195,294.97	19,900,969.10	(3,705,674.13)	23%	
12	17,806,452.37	28,293,935.05	(10,487,482.68)	59%	
13	12,220,832.23	8,741,575.20	3,479,257.03	28%	
14	11,387,584.90	11,098,716.14	288,868.76	3%	
15	4,973,374.71	6,530,037.43	(1,556,662.72)	31%	
16	15,613,615.70	18,764,049.49	(3,150,433.79)	20%	
17	3,791,727.89	3,052,137.92	739,589.97	20%	
18	19,919,358.50	21,324,631.45	(1,405,272.95)	7%	
19	29,248,493.16	19,888,306.82	9,360,186.34	32%	
20	2,893,236.56	2,598,438.09	294,798.47	10%	42%
21	22,589,408.08	18,765,601.08	3,823,807.00	17%	
22	23,986,009.78	21,474,081.67	2,511,928.11	10%	
23	17,079,813.18	17,758,228.92	(678,415.74)	4%	
24	8,126,358.30	10,664,689.10	(2,538,330.80)	31%	
25	2,676,461.89	8,376,170.92	(5,699,709.03)	213%	
26	3,804,284.70	5,699,843.62	(1,895,558.92)	50%	
27	6,488,621.01	8,574,459.91	(2,085,838.90)	32%	
28	2,911,299.76	807,755.40	2,103,544.36	72%	
29	12,478,276.57	23,942,817.00	(11,464,540.43)	92%	
30	7,981,174.18	2,620,815.27	5,360,358.91	67%	
31	11,783,548.53	23,818,789.10	(12,035,240.57)	102%	
32	27,657,723.39	18,982,397.61	8,675,325.78	31%	
33	8,119,134.94	4,604,431.86	3,514,703.07	43%	
34	15,173,783.08	20,256,849.65	(5,083,066.57)	33%	
35	24,788,888.32	20,196,877.44	4,592,010.88	19%	
36	22,296,477.64	13,293,962.04	9,002,515.60	40%	
37	4,308,488.64	4,383,572.78	(75,084.15)	2%	
38	29,877,450.25	14,293,742.85	15,583,707.40	52%	
39	8,270,445.47	8,966,230.42	(695,784.95)	8%	