A Model for Forecasting Owner's Project Management Resources

by

Hady Elkholosy

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in
Construction Engineering and Management

Department of Civil and Environmental Engineering
University of Alberta

Abstract

This research involves the development of a forecasting model that will aid project managers in estimating the owner's project management staff requirements for a construction project utilizing historical data. This will support project managers in the decisions of hiring new resources or working with the current available ones as well. The objective of this research is to provide a methodology that will help in forecasting the project management staff hours for a given project by studying the factors that impact staff requirements and developing an analytical model utilizing data mining techniques. The research includes understanding industry practices in resource allocation and estimation, collecting and analyzing historical data, and evaluating different machine learning algorithms, such as artificial neural networks, multiple linear regression, KNN and random forests, to forecast project management hours.

The contributions of this research are combining industry practices and literature review to identify the projects' features that affect the staffing requirements, proposing a data acquisition model that will help industry practitioners in collecting these attributes properly, and developing a neural network model to utilize the data collected and forecast the owner's project management hours required for buildings, infrastructure and industrial projects.

Acknowledgements

I would like to thank my graduate studies supervisors Dr. Ahmed Hammad and Dr. Simaan M. Abourizk for their guidance and support throughout this research. Appreciation is extended to Dr. Ramzi Roy Labban for his assistance and to our industry partners whom I would not have completed my work without their help and feedback.

Many thanks to Marial Al-Hussein and Mickey Richards for their contribution regarding the application development and manuscript editing, respectively.

Finally, without the support of my lovely family, this could not have been possible. Thank you to my loving wife, who kept pushing me throughout this process. To my mother, father and sister who have been my biggest supporters throughout my life and believing in me more than I ever have myself.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ANN	Artificial Neural Network
BP	Back Propagation
BIM	Building Information Modeling
CM	Construction Management
CBR	Case-based Reasoning
DBB	Design Bid Build
DB	Design Build
DSS	Decision Support System
ERP	Enterprise Resource Planning
ERD	Entity Relationship Diagram
IBC	International Building Code
KDD	Knowledge Discovery in Database
KNN	*k*-Nearest Neighbor
LHRs	Labor Hours
M	Million
MAPE	Mean Absolute Percentage Error
MAE	Mean Absolute Error
MLP	Multi Layer Perceptron
PM	Project Management
PDM	Project Delivery Method
PDF	Probability Density Function
WSS	Within-Cluster Sum of Square

Chapter 1: Introduction

1.1 Overview

Construction projects are associated with different features and objectives, which makes each project unique in aspects such as the resources required and execution time. Since the construction industry is a labor-intensive industry, allocating the required staff to an upcoming project is a challenging task. Accurate labor estimates are key to achieving successful construction projects. One labor category that leads to better performance in projects is project management (PM) staff.

PM staff are considered indirect resources who oversee construction projects from start to finish by planning, scheduling, controlling and evaluating the project to reflect its budget and schedule. Tasks carried out by PM staff and the effort needed by PM individuals involved in a project differ between the owner side and the contractor side. However, both owner and contractor PM staff are essential for the project success as they view the project from different perspectives. This research focuses on estimating the owner's PM staff requirements for a construction project. The PM staff includes different resource categories that can divided into program managers, project managers, project coordinators, and project assistants.

Currently, manpower estimates are accomplished by project managers utilizing a mixture of both analytical analysis and individual's judgement. The knowledge and the experience of PM staff play important roles in estimating the required resources for an upcoming project. However, this method occasionally provides unreliable and inaccurate estimates because the estimators' experiences vary and each project is unique. Also, data collection may be done inconsistently, and several important features that need to be collected may be overlooked. This often causes over- or underutilization of resources, and the workload becomes unfairly divided among project staff. The

project individuals could end up stressed, underperforming their assigned work, or quitting their jobs due to the workload. Thus, the goal of this research is to develop a proper framework for data collection and a manpower forecasting model to provide owners with more reliable estimates for the PM staff required for a construction project, resource availability, and staffing requirements decisions.

1.2 Research Objectives

This research attempted to develop a model that will forecast owners PM resource hours for a given project assignment. The research objective was accomplished by achieving the following:

1. Investigating industry practices and literature to identify key project attributes impacting PM resources requirements.

2. Collecting historical project data from different owners to use in developing the forecasting model.

3. Evaluating different machine learning algorithms and their capacity in predicting the PM hours required for a given project.

1.3 Research Methodology

To achieve the research objectives, the research methodology investigated the factors impacting the manpower requirements, collected historical projects data, applied data mining techniques to obtain useful knowledge from the data, and received experts' feedback on the developed forecasting model.

The first stage of the research project involved carrying out an extensive literature review of methods for forecasting manpower demands and workload and assessing their applicability to the

defined problem. The literature review identified important project features for estimating manpower requirements. The second stage involved understanding current practices in the construction industry to estimate and allocate PM staff, including factors considered by the project managers during the process of assigning PM staff to a new project. The third stage involved collecting historical data that included the projects' size, complexity, duration and labor hours spent by the PM staff on these projects. This stage also involved cleaning and preparing the data to be used in the next stage. The fourth stage involved developing the forecasting model, including the application of data mining techniques to identify the most significant project features impacting the PM hours. Specific different machine learning algorithms were evaluated, such as multiple linear regression, random forests, *k*-nearest neighbor (KNN) and artificial neural networks (ANN). The fifth stage included development of the decision support system that implemented the findings from the previous stages. The model consists of two main modules. The first module is the data acquisition system that can track and store project information. The second is the analytical model that has the ability to forecast PM hours required for a new project utilizing the collected data.

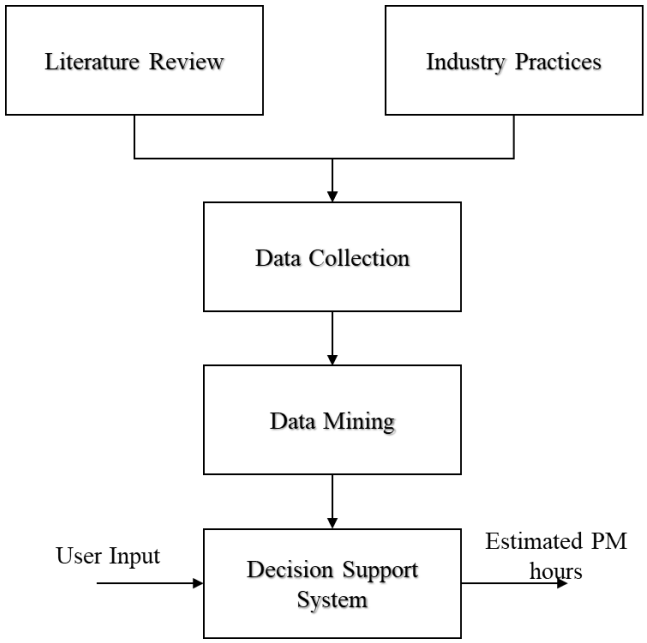


Figure 1 Research Methodology

1.4 Expected Contributions

The contributions of this research can be classified into academic and industrial contributions.

The academic contributions of this research are:

1. Combining both literature review and industry practices in identifying the features impacting PM staff requirements.

2. Applying machine learning algorithms to forecast the PM staff requirements.

The industrial contributions are:

1. Developing a data acquisition system for tracking and storing projects data to predict PM hours for future projects.

2. A methodology for forecasting PM hours required for a project using an artificial neural network model.

1.5 Thesis Organization

Chapter 2 presents the literature review for this research. The literature review chapter focuses on the approaches proposed for forecasting manpower demand and data mining techniques used for prediction. Chapter 3 introduces the methodology adopted in this research, the data collection process, and the proposed data acquisition model. Chapter 4 outlines the procedure of developing the forecasting model. Chapter 5 presents the developed application that consists of two modules – the database that stores all projects information and the analytical model that carries out the forecasting. Chapter 6 includes the final discussion and proposed future research.

Chapter 2: Literature Review

2.1 Introduction

The literature review chapter focuses on three main topics: labor resources management, manpower forecasting models, and types of machine learning algorithms in data mining. Labor resource management is investigated to understand the process of resource allocation and identify the factors that are considered by project managers during this process. The manpower forecasting models are studied to discover the techniques adopted by researches to estimate the required resources. Data mining techniques are explored due to their capabilities in discovering useful knowledge in data and for their capacities in performing predictive analytics. The main objectives of this chapter are to identify construction projects features affecting the total PM hours required, any gaps in the literature, and applications used in forecasting project staff.

2.2 Labor Resource Management

Labor resources are crucial to companies and projects, as their skills and capabilities support the success of construction projects (Chen et al., 2008; Dabirian et al., 2019; Fini et al., 2018). The resource allocation process in a multi-project environment tackles some challenges, such as assigning the proper individuals and optimal staffing levels (Engwall & Jerbrant, 2003; Hendriks et al., 1999; Silva & Costa, 2013). Therefore, experts' knowledge and experience is required to assign PM staff to finish the project within budget and schedule. The resource allocation process becomes more complex as the number of projects increase because the number of available personnel and their knowledge are limited and vital within an organization (Hendriks et al., 1999; Silva & Costa, 2013). Frequently, PM staff end up being over- or underutilized, and their workload becomes unfairly divided, causing stress or anxiety and impacting performance (Leung et al., 2008).

Fini et al. (2018) proposed a methodology for optimizing employment and firing decisions in a construction organization using a dynamic programming algorithm. This framework considered human resource strategies, workers' rights, and technological and managerial aspects of a project in the optimization process. The information was gathered from building information modelling (BIM), a project performance database, and contract documents. This method improved the estimation accuracy of labor resources by using BIM; however, the estimates were on a macro level. Furthermore, the proposed approach did not consider the market variability and employee turnover.

A methodology based on dynamic programming was developed by Silva and Costa (2013) to assist in allocating labor resources to software projects. The methodology considered different aspects of the project and staff capabilities. Its main objective was to minimize the time required to finish the project by determining the number of professionals and identifying who is the best fit. The framework considered the complexity of each project, current staff capabilities, and the skills required to finish the project. This study disregarded the possibility of having one staff member working on multiple projects, as the author stated that it will lower overall efficiency which usually is not the case in the real world.

Ballesteros-Pérez et al. (2012) suggested that a team's performance relies heavily on the members' social ties and communication. Accordingly, they proposed a framework to decide on individuals who can work together and form a good team to maximize work efficiency and achieve the best performance. The framework was based on sociometry, a technique developed by Jacob Levy Moreno (1960) that measures the interpersonal relationship between two individuals or a group of people. This study used quantitative sociometric tests to evaluate the group cohesion, then mathematical calculations were adopted to measure the group efficiency. Finally, a personal

contribution index of a member was calculated using *Individual Sociometric Indices*. This decision support system provided project managers information required to build a team with good interpersonal relationships, which, therefore, should perform better. However, team members should meet the required technical skills and experience prior to applying this sociometric approach.

Jun and El-Rayes (2011) developed a multi-objective optimization model to optimize resource leveling and allocation in construction projects. This model adopted a multi-objective genetic algorithm that provided construction planners with some features, including determining and minimizing the undesired fluctuations, shortening project duration, and developing optimal trade-offs between resource utilization efficiency and project duration.

André et al. (2011) outlined two main issues that affect project success: improper resource assignment and conflicts among the project team. Usually, resource allocation and team formation are based on the manager's experience, as well as workers' constraints and skills; however, these data are often not documented in a database for future use. This research utilized the Delphi technique to identify the key factors that should be considered in the staff assignment process. The surveys showed that project complexity, project importance and risk, team collaboration, members' competence and availability all need to be considered in staff assignments. Moreover, the experts agreed that the following points impact the assignment factors: assigning an individual to multiple projects, allocation to other projects depending on worker's role in the previously assigned projects, personnel having more than one role in a project, and availability of unsuitable roles. Furthermore, data mining techniques were applied to investigate psychological tests and formulate rules for building the project team. Although, these rules were helpful, they could not classify engineers' competence, and the results cannot ensure 100% work performance quality.

Lin (2011) developed a decision-making model for allocating PM staff in remote construction projects that considered total project costs. There are two approaches that construction firms often adopt: assigning from the current available staff or hiring a local temporary one. Project losses were projected by assessing the competence of in-house and local personnel. Analytical hierarchy approach was used in the evaluation and weighting process of staff competence. This study concluded that the in-house PM staff are preferable to local ones, unlike the site engineers. Although, this framework accounted for several factors, it overlooked a few other aspects such as market growth and workforce training.

Ahadzie et al. (2008) proposed a competency-based multidimensional model that indicated project managers' performance behaviors and outcomes throughout the project lifecycle for mass house building projects. Stepwise selection was adopted to identify the most significant variables for predicting the project manager's performance in the construction phase. This study suggested that the following factors hugely impacted the model performance: familiarity with site layout techniques for repetitive tasks, commitment to help subcontractors to finish the project on schedule, knowledge of proper technology for improving performance on repetitive tasks, time management on house units, ability to resolve conflicts and preserve good relationships, and flexibility when approached by subcontractors to resolve their difficulties. The presented model had a MAPE of 5% and 0.69 R^2, which are good; however, the model was limited to mass house building projects and a relatively small sample size.

Wu and Sun (2006) introduced a non-linear programming and genetic algorithm model for project scheduling and manpower allocation. The model was built to overcome the challenges facing the multi-project environment, which involves reallocation of staff to different activities of concurrent projects. The model's objective function was to minimize subcontracting costs. The authors

disregarded the scatter effect, which represents the decrease in efficiency when the number of individuals working on a task increase.

In a multi-project environment, there is often ongoing competition between different project managers in terms of priorities, staff, and resources. Engwall and Jerbrant (2003) carried out two qualitative case studies for two multi-project organizations to generate concepts, theoretical models, and practical issues. The authors outlined that the resource allocation process in a multi-project organization is more complex than has been previously mentioned in the literature, as it included politics, knowledge and common sense. Finally, multi-project management should start building motivation and accounting systems, rather than focusing solely on resource allocation.

Hendriks et al. (1999) focused their study on optimizing the resource assignment process in a large R&D organization by identifying five elements that were critical for the improvement of this process. These elements were long-term resource allocation (focused on the business plan of each discipline for the upcoming year at least), medium-term resource allocation (focused on day-to-day planning for reviewing the project portfolio at least once a year), short-term resource allocation (the main input for day-to-day planning of resources for the near-term weeks), links between the allocation processes, and continuous feedback. Also, the authors introduced two indicators, project scatter factor (representing the number of individuals required to complete a one-year task) and resource dedication profile, that could simplify the implementation of the medium-term resource allocation process and improve the project's results significantly.

2.3 Manpower Resource Forecasting Models

Forecasting workforce requirements is a critical and strategic managerial practice for the public and private sectors due to the huge influence of labor resources costs on the financial assets of an

organization (Meehan & Ahmed, 1990; Sing et al., 2014). Additionally, manpower forecasting plays an important role in labor resource planning, which is essential for the business planning process (Meehan & Ahmed, 1990; Wong et al., 2007). Sustaining staff levels in a construction firm is a challenge due to the temporary nature of projects (Fini et al., 2018). Also, labor requirements fluctuate during project execution due to the distinctive nature of each phase in a construction project. Accordingly, forecasting accurate staff requirements is difficult because of substantial variations in output (Sing et al., 2012). Thus, the selection of a proper forecasting technique is critical to achieve reliable results with good accuracy (Wong et al., 2011). A number of studies had been conducted to forecast staffing demand and allocate labor resources properly on a market level, while the others tried to predict required staff and allocate them on a project level, using various approaches.

2.3.1 Forecasting Models for Market Demands

Sing et al. (2014) adopted the triangulated approach to develop a robust manpower forecasting model that could predict the technicians required in the construction industry in Hong Kong. The model constructed a qualitative model through structured interviews with construction companies' senior management and developed a quantitative model based on the outcomes of the qualitative model. This model provided construction employers in Hong Kong with the knowledge to determine the future workforce demands and develop trainings' strategies.

Sing et al. (2012) used distributed lag model and labor multiplier approach to predict the labor demand in different economic conditions. It was assumed that there was a constant relationship between construction output and resources required per construction unit in the short-term. Using this correlation, a fixed labor multiplier was determined for different trades. The model provided decision makers with knowledge about economic impacts on labor demand. This study did not

consider the variability of labor demand during the projects and assumed that these multipliers were constant. Moreover, the selection of a proper economic scenario is crucial for accurate results.

Ho (2010) developed a gray model that predicts the construction manpower demand based on a limited amount of data, unlike the other models that required a large amount of input data. A first-order gray model includes three main procedures: accumulated generation to identify any laws behind the raw data, using the accumulated generation to develop first-order gray differential equation, and inverse accumulated generation to restore the data to its original form. The model could forecast the national manpower demand for one quarter ahead using limited data with an acceptable accuracy. Also, it could be used in forecasting manpower demand by occupation or region, if the available time series sample was not less than four points. The first-order gray model that was used is one of the most basic models, where a number of more complicated models might provide better results.

A study was carried out by Wong et al. (2007) to establish a relationship between construction labor demand and a set of interconnected economic factors, such as construction output, worker pay, material costs, interest rate, and labor productivity. Then, a vector error correction (VEC) model was developed to forecast the aggregate manpower demand. The model was robust and determined a relationship between labor demand and the above-mentioned economic factors. Decision makers and training planners can utilize the results of the forecasting model to determine the manpower resource requirements and plan for training. Even though, the model could predict medium-term cumulative labor demand, it required further development to account for industrial advances and workload in order to forecast market demand by occupation.

2.3.2 Forecasting Models for Project Requirements

Dabirian et al. (2019) developed a dynamic model that is capable of estimating labor need for a construction activity and effectively allocating the workforce to the project. A system dynamics approach was used in modeling and evaluating the different strategies of workforce allocation. Different categories were incorporated that involved activity, productivity, quality, and workforce; and under each category different factors were considered. The model proposed a solution to overcome fluctuations in staffing needs and difficulties from staff turnover. Although, the learning effect of unskilled labor was reflected by the authors, other human factors such as welfare and motivation were overlooked.

Yang and Kim (2019) carried out a study that showed a strong relationship between the number of site staff, classified by work type, and the project's technical risks such as site conditions, construction methods, subcontracting conflicts, and permits. The allocation methodology proposed in this study improved the accuracy of estimating the site engineers for a project, though the regression models provided could not be generalized because construction firms adopt different techniques in manpower estimates, risk identification, and evaluation processes. Moreover, the regression model can be applied only to apartment building projects and the dataset that was used for building and testing the model was relatively small, including only 31 projects.

Hammad et al. (2014) proposed a methodology that used labor resources data from industrial construction projects and knowledge discovery in database (KDD). The model was able to forecast working hours and duration from the historical data through data mining techniques. The model improved the process of estimating labor for future industrial projects. The error from the model was approximately 25%, which is a bit high; however, it was better than the accuracy of the current

estimate practices. Also, there was room for improvement by generalizing the model to work with different types of projects not only industrial projects.

Othman et al. (2011) developed a nonlinear programming model to reduce the employment, firing, training, and overtime costs. The model can estimate the number of required staff for each work type, trained employees, and overtime hours. Human factors such as breaks, vacations, and training periods were considered in the estimation process. A decision support system (DSS) was then established to help in the workforce planning process and facilitate the interaction between workers, managers and systems. The model disregarded other human factors such as: the learning effect, experience, and motivation, all of which might improve the forecasting accuracy.

Wong et al. (2008) developed forecasting models to predict total workforce requirements and 10 important trade demands for construction activities. This research introduced additional factors, other than the project size and type, to be considered in the forecasting models such as: construction method, degree of mechanization, project complexity, management attributes, and expenditures on electrical and mechanical services. Multiple linear regression models and stepwise selection techniques were adopted to identify the relationships between the variables and pick the statistically significant ones. This study concluded that the construction cost was the most significant factor in determining resource requirements. However, project complexity, site conditions, and project type should not be overlooked as they affected the total labor demand. Moreover, trades requirements were defined using site conditions, participation of construction knowledge during design, and site coordination. The dataset used for developing the model was relatively small (50 projects), and the model was tested using only four projects where overfitting could have occurred. Another limitation was that the model required continuous review and

updating to incorporate any forthcoming fluctuations in design, technology, and construction methods.

Chen et al. (2008) developed a model that forecast the onsite supervisory staff and its costs using mathematical analysis and case-based reasoning (CBR). The authors identified eight important variables for analysis: project type, owner type, structure type, contract price, floor area, structure height, location, and project duration. The proposed model had an accuracy rate of 88.47% for predicting the onsite supervisory costs and staff allocation. It concluded that crashing schedules and floor area have a significant impact on the results. The model was limited 65 projects, most of which had a relatively small area and un-crashed schedules that might have affected the calculations behind the CBR model.

Bell and Brandenburg (2003) developed a tool that can predict staff requirements for the construction project management team in transportation projects. This research assumed that the project's type and estimated total cost are the main factors in determining the manpower demand. A regression analysis showed an exponential relationship between project cost and spent man-hours for the project categories including bridges and overpasses, resurfacing, rehabilitation, etc. This study mainly focused on obtaining useful information from historical data which made it difficult to predict the overall manpower required due to the lack of data available in some of the categories. Furthermore, the presence of outliers in the generated regression analyses without further investigation reduced the accuracy of the prediction tool.

A quick and reliable approach was developed by Proverbs et al. (1999) to predict labor requirements and costs for construction projects during the initiation stage where the design information is premium. This method utilized productivity rates adopted by contractors' planning

engineers for a typical high-rise concrete building to generate a “Labor Estimate Factor”. This factor represents the man-hours required per square meter of floor area. Then, the national wage rates for the different countries were used to forecast the labor costs. This method was limited to in-situ concrete buildings in three countries (France, Germany, and UK); thus, it needs further developments to fit any type of structure and location.

Persad et al. (1995) carried out a study to forecast engineering staff requirements for highway preconstruction activities utilizing historical data for completed projects. A simple linear regression was calculated for the construction cost, engineering man-hours, and engineering cost to detect the relationships between these aspects. Furthermore, multiple regression with stepwise selection of significant variables was performed, and the significance of project type was examined. This research concluded that the project construction cost and the project type performed well in forecasting engineering staff requirements. Also, it was found that the engineering cost was highly affected by project complexity; however, no factors other than the project type were considered. The model used the estimated construction cost to determine the required manpower, which continuously fluctuates until the contract is signed.

Meehan and Ahmed (1990) proposed a model to predict categorized staff requirements for an electrical utility company. The demand models were developed using regression analysis that considered certain workplace and staff conditions specific for this entity. The models were based on some dependent variables, including total staff required, professionals, and managers, and independent variables involving capital budget, sales revenues, and assets. The models provided good quality results that applies only to this firm and cannot be generalized.

2.4 Data Mining Techniques

Bhatia (2019) defined data mining in his book *Data Mining and Data Warehousing: Principles and Practical Techniques* as "a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large databases. The patterns must be actionable so they may be used in an enterprise's decision making." This is one of the most recent definitions of data mining that covers its key elements in a simple way. Data mining has been widely used by different industries including information technology, construction, sales, banking, etc. to find valuable information in historical collected data that could enhance their business outcome in the market. This is also known as KDD.

There are two types of classification:

1. Posteriori classification: A supervised machine learning approach where the class label of each training tuple is predefined. The prediction of a class label is achieved by learning a mapping or function that is denoted in the form of decision trees, classification rules, or mathematical formulae (Bhatia, 2019; Han et al., 2012; Witten et al., 2011).

2. Priori classification: An unsupervised machine learning approach, where the target class of each training tuple is not provided. This requires the use of clustering to divide the dataset into different groups where the objects of a group or cluster have similar characteristics. Then, each cluster can be labelled will enable the classification process to take place to make predictions (Bhatia, 2019; Han et al., 2012).

Due to the high capabilities of machine learning algorithms in predictive analysis, data mining techniques have been adopted to solve defined problems using historical projects data. The

obtained knowledge can help in estimating owner's project management hours requirements for an upcoming project.

2.4.1 Decision Trees

The decision tree classifier makes predictions based on a set of "if…then…" conditions that are similar to if statements in numerous programming languages (Bhatia, 2019). The tree consists of a root node, the topmost node, that is split into branches based on all possible values, then this split is repeated for each branch until reaching a node where all instances have the same classification (Witten et al., 2011). This data mining technique is widely used in medical diagnosis, data structures, botanical classification, and psychology; however, it is not commonly used in the pattern recognition field from statistical point of view (Gorunescu, 2011).

There are three traditional approaches for decision trees (Gorunescu, 2011):

1. Classification trees: The prediction result is the labelled class of the data.
2. Regression trees: The prediction result is a real number.
3. Classification and Regression Tree (CART or C&RT): The prediction result is a combination of the above situations.

The main advantages of the decision tree classifier lie in the simplicity of the generated rules, domain knowledge is unnecessary, learning and classification processes are easy and quick, numerical and categorical data are not restricted, and building costs are low (Bhatia, 2019; Gorunescu, 2011). Despite this, there are some drawbacks with decision trees as they are complex and time consuming when dealing with large datasets. Also, any minor changes in the data can

affect the entire model, and sometimes their prediction accuracies can be relatively low, resulting in huge classification costs (Bhatia, 2019; Gorunescu, 2011).

Decision trees have been used in several applications in the construction industry, particularly safety, contract administration, and risk management (Amiri et al., 2016; Dey, 2012; Lee et al., 2004).

2.4.1.1 Random Forests

Random forests are considered as bagging ensemble methods that consists of a set of decision trees (Bonaccorso, 2018; Raschka, 2017). The main objective of random forests is to develop a more robust model that has a better performance and is less vulnerable to overfitting through averaging several decision trees that individually suffer from high variance (Bonaccorso, 2018). The splitting strategy in random forest is built on a medium level of randomness differing from a single decision tree (Raschka, 2017). A random subset of features in each tree is used to get the threshold that best splits the data. Thus, there will be several trees that are weekly trained, and each one will provide a different prediction (Raschka, 2017). Simultaneously, each tree will have the best performance in a portion of the sample space and provide inaccurate estimates in other portions (Raschka, 2017). There are two methods used to calculate the predictions of the trees: the majority vote and averaging the results; both methods often provide very similar results (Raschka, 2017). The method used in this research is the probabilistic averaging which is adopted by Scikit Learn library in Python (Pedregosa et al., 2011).

One of the main advantages of random forest is that choosing good hyperparameter values is effortless compared to ANN. The number of trees (k value) is the main parameter that needs to be emphasized, because the larger the number of trees, the better performance the random forest will

have (Bonaccorso, 2018; Raschka, 2017). Usually, a value of *k* between 20 and 30 trees provides better results than a single decision tree, as increasing that number will impact the computational costs (Bonaccorso, 2018; Raschka, 2017).

Random forests have been used by researchers to solve various problems in the construction industry such as detecting corporate's misconduct, predicting occupational accidents and BIM labor costs (Huang & Hsieh, 2020; Kang & Ryu, 2019; Wang et al., 2020).

2.4.2 Linear Regression

Linear regression is often used when the class and all the attributes are numeric, as it is an excellent technique in numeric predictions (Witten et al., 2011). The regression term is used by statisticians to represent the process of forecasting a numeric amount (Witten et al., 2011). The main purpose of this method is to predict the value of a certain variable based on the values of other variables, with predefined weights determined from the training data, and represented in an equation form describing the relationships among the dataset (Gorunescu, 2011; Witten et al., 2011). The regression equation is determined by the Least Squares Method that naturally minimizes the distance between the data points and the points on the developed regression line (Gorunescu, 2011).

$$x = w_0 + w_1 a_1 + w_2 a_2 \ldots + w_n a_n \quad (1)$$

Where:

- x is the output (desired variable's value)
- $a_1, a_2, \ldots, a_n$ are the attributes' values
- $w_0, w_1, \ldots, w_n$ are the weights.

The main disadvantage of this method is the linearity; when the data show a nonlinear dependency, a poor fit may result (Witten et al., 2011).

Linear regression models have been widely used by many researchers in the construction industry to forecast manpower demand, construction costs, and buildings waste and to help in design decisions (Holmes et al., 2019; Lowe et al., 2006; Parisi Kern et al., 2015; Wong et al., 2011).

2.4.3 Artificial Neural Networks (ANNs)

ANNs provide methods to model large, complex problems that involve many interrelated variables (Mourya & Gupta, 2012). They have been employed in several areas such as modeling, time series analysis, pattern recognition, signal processing, etc. (Gorunescu, 2011). ANNs tolerate noisy data, are able to predict without training, and can classify without knowledge of the attributes' relationships (Gorunescu, 2011; Han et al., 2012). The main disadvantages of ANNs are that they require prolonged training time, knowledge, and experience to determine their parameters, and it can be difficult to understand the meaning behind the calculated weights of the nodes (Han et al., 2012). ANNs can be used in both classification and regression problems (Mourya & Gupta, 2012).

ANNs were built to process information in a manner similar to the human brain (Gorunescu, 2011). They consist of a set of interconnected input and output units where each link has a weight associated with it (Han et al., 2012). To be able to predict the output of the input tuple, the network learns by modifying the weights (Han et al., 2012). There are different types of neural networks that are used for different purposes, but for the defined problem, this research will focus on multi-layer feed forward and multi-layer perceptron (MLP) types in which the backpropagation algorithm is applied.

A multi-layer feed forward ANN is composed of an input layer, one or more hidden layers, and an output layer (Han et al., 2012). Each layer consists of a number of units. The input units in the input layer represent the attributes used in the training model (Han et al., 2012; Mourya & Gupta, 2012). After the input passes through the input layer, they are fed into the second layer (hidden layer) inputs with assigned weights (Han et al., 2012; Mourya & Gupta, 2012). The output of the hidden layer can be fed into multiple hidden layers or fed directly to the output units in the hidden layer (Han et al., 2012). The output layer, which represents the prediction outcome, is made from the calculated outputs of the last hidden layer (Han et al., 2012; Mourya & Gupta, 2012).

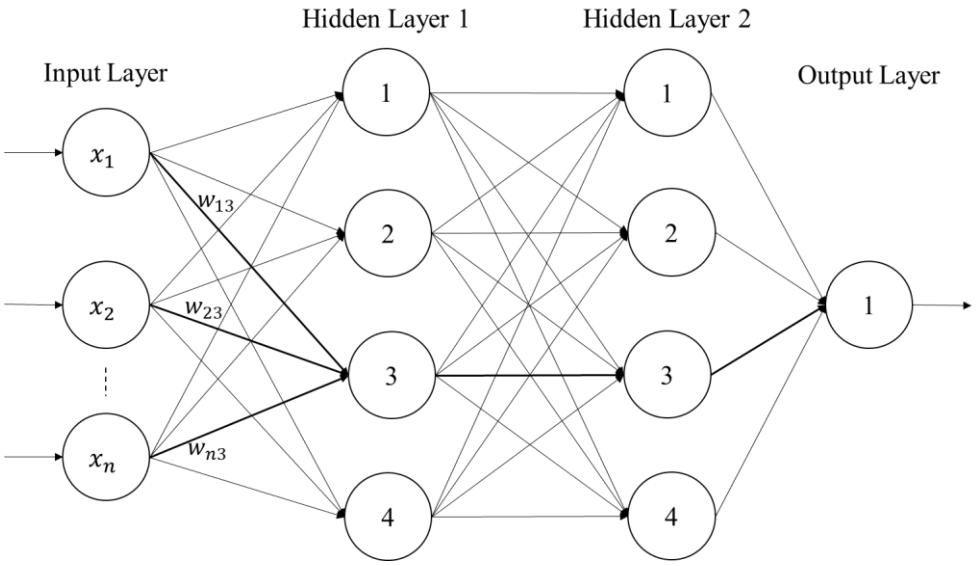


Figure 2 Multilayer feed forward neural network

A multi-layer feedforward ANN can also be named MLP ANN as they both have the same network structure. Theoretically, an MLP with just 2 hidden layers is capable of solving any practical problem (Gorunescu, 2011). As a result, it is the most common type used in software packages such as Scikit Learn library in Python (Pedregosa et al., 2011).

Back-propagation (BP) algorithm is one the most common learning techniques for different types of ANNs. BP stands for *backwards propagation of errors* (Gorunescu, 2011). The concept behind

BP is minimizing the error between the actual target value and the predicted output through an iterative process for the training set (Gorunescu, 2011; Han et al., 2012; Mourya & Gupta, 2012). The calculated error is based on a predetermined error function *E*, which is a function of nodes' weights ($w_1, w_2, \ldots, w_p$) where:

$$E = E(w_1, w_2, \ldots, w_p) \quad (2)$$

After obtaining the result, the algorithm moves backward to assign the computed error to the nodes in each hidden layer based on their weights (Gorunescu, 2011; Han et al., 2012; Mourya & Gupta, 2012). Then, the weights at each node in the hidden layers are modified to achieve the least error (Han et al., 2012; Mourya & Gupta, 2012).

ANNs have been used in various complex problems in the construction industry for almost three decades. They have been used to rate subcontractors, analyze construction claims, forecast project cost in the preconstruction phase, and estimate labor productivity (Ahiaga Dagbui & Smith, 2014; Art Chaovalitwongse et al., 2012; M. Lu et al., 2000).

2.4.4 Bootstrapping Neural Networks

An ANN is a nonlinear statistical method that can be used for classification or regression problems in various industries such as manufacturing, construction, financial, and medical (Papadopoulos et al., 2001). The network reliability and usability is important to have a generalized model, and they can be increased when the predictions are supported with confidence intervals (Franke & Neumann, 2000; Papadopoulos et al., 2001). ANN ensemble techniques have been an interest for some researchers in which multiple neural network estimates are combined to provide performance better than the one provided by a single network (Carney et al., 1999; Papadopoulos et al., 2001; Zhang, 1999). Bootstrapping is one of the popular methods that are used for quantifying the

standard error of a statistical parameter as it is a well tested method in linear and non-linear statistics areas (Carney et al., 1999; Franke & Neumann, 2000; Paass, 1992). Also, it can be used to calculate confidence intervals for predictions by combining multiple neural networks predictions to improve model accuracy and robustness (Carney et al., 1999; Franke & Neumann, 2000; Paass, 1992; Zhang, 1999). The main concept of the bootstrap method is resampling the available dataset with replacement and training a single network on each of the bootstrap samples (resampled instances of the original dataset) (Zhang, 1999).

Breiman (1996) outlined that bootstrapping works well for unstable machine learning algorithms such as neural networks, classification, and regression trees. On the other hand, bagging will not work properly with stable procedures such as *k*-nearest neighbor. In this research, both ANN and random forest are used to forecast the owner's PM hours for a given project. The random forest algorithm already adopts bootstrapping with replacement to make a prediction through averaging the result of multiple decision trees. Alternatively, implementing neural network bootstrap ensembles needs to be done in two different phases. First, building a neural network model to be used in making one prediction. Second, creating bootstrap samples from the original dataset and train the neural network model using the bootstrap samples to get multiple predictions.

Using neural network with bootstrapping will help in estimating confidence and prediction intervals, which will improve the uncertainty of the estimates and the robustness of the model (Carney et al., 1999; Paass, 1992; Zhang, 1999). This technique has been used in various areas such as computer science, economics, industrial, energy and construction. It was used to provide range prediction for construction costs, nuclear transient processes, industrial equipment degradation, option pricing for stocks and building robust nonlinear models (Lajbcygier & Connor, 1997; Lins et al., 2015; Sonmez, 2011; Zhang, 1999; Zio, 2006).

2.5 State-of-the-Art Discussion

Forecasting staff requirements has been a topic of interest for researchers in different fields. The developed manpower forecasting models which are previously mentioned can be divided into three main categories. The first category comprises estimating the total manpower demand and trades on a market level based on economic factors. The second involves forecasting the labor demand on a project level using different project characteristics that include technical and managerial aspects. The third category involves estimating the required labor for a construction activity considering both activity and workforce attributes.

Few solutions have been offered to forecast the direct or indirect staff required for an upcoming project. Both quantitative and qualitative approaches have been used to identify the most significant project attributes impacting the number of the required staff (Bell & Brandenburg, 2003; Chen et al., 2008; Wong et al., 2008). Although, several factors were considered important based on experts' opinions, they were disregarded in the models developed by the researchers. The proposed models were robust; however, few of them were developed considering less than 50 construction projects, which often provides optimistic results. Moreover, the proposed models were limited to a specific company or certain types of projects and cannot be generalized to other types of projects or fields. Also, most of the research was carried out from the contractor's or the consultant's point of view rather than the owner's perspective. Overall, none of the provided solutions have the scope or the accuracy to be widely accepted.

The key investigation of the research in this thesis is to try to combine the current practices and literature to forecast the owner's PM staff required for any type of construction project. The purpose is to have a generic model that provides project managers with reliable results. This is achieved by proposing a data acquisition model to help project managers in collecting the project

information in a structured way which will offer a better forecasting accuracy, developing a forecasting model utilizing machine learning algorithms to forecast the hours required, and developing a decision support system for owners to help in the allocation process and support in hiring/staffing decisions.

Several proposed models used linear regression analysis, while others used case-based reasoning and non-linear programming to estimate the required manpower. This research used multiple machine learning algorithms, including linear regression, random forests, and ANNs, due to the capacity of data mining in providing good estimates and predictions that can then improve the business. Then the performance of machine learning algorithms was compared to determine which algorithm fits the defined problem and has better prediction accuracy.

Chapter 3: Methodology for Forecasting PM Staff Requirements

3.1 Introduction

The objective of this thesis is to forecast the required PM staff for an upcoming project using historical project data and to develop a framework for collecting key project factors that will help in forecasting the total hours. The proposed methodology combines the literature with experts' knowledge to develop a reliable forecasting model that can be used by practitioners, as shown in Figure 3. This framework is comprised of multiple phases, including understanding industry practices and methodologies, investigating other researchers' approaches in forecasting manpower demand on a project level, and determining the key factors impacting the total staff required. The next step is to propose a data acquisition model for collecting project information properly for forecasting PM staff requirements. Then, historical project data were collected based on the findings of the literature review and industry practices. After data cleaning and transformation, machine learning algorithms were adopted to identify the significant project features and develop a model that can forecast staff requirements. Finally, the model was validated the industry experts.

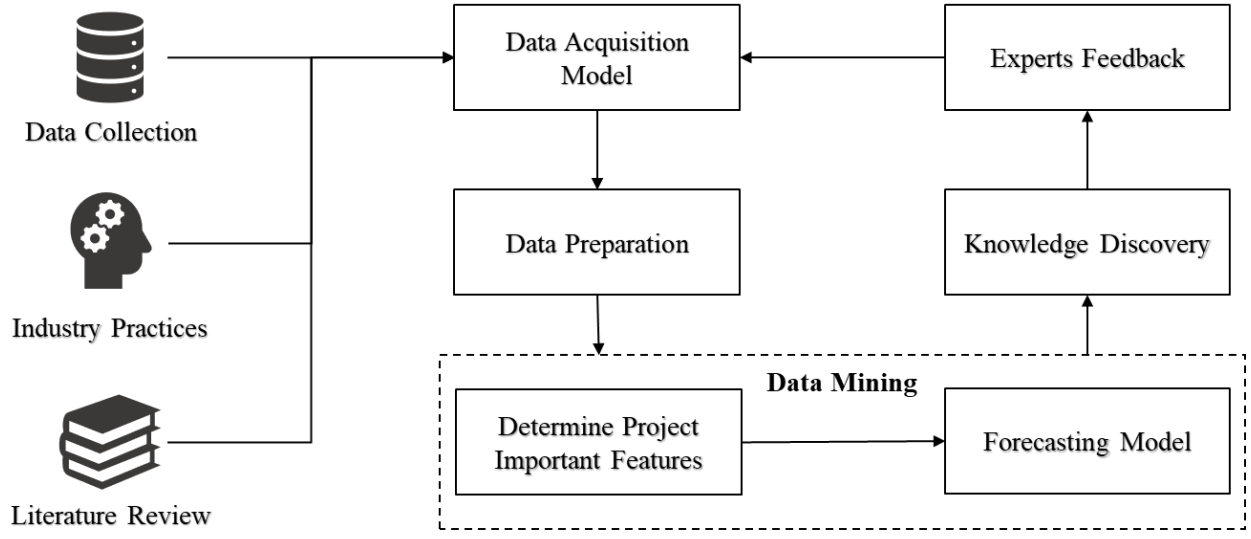


Figure 3 Methodology of Developing Data Acquisition System and Forecasting Model

The objective is to discover which projects features impact the total PM hours required for a new project. Also, the developed forecasting models will allow us to find the weights of these factors and interpret the results. The outcome of this study needs to be reasonable and understandable to project managers, so it can support them in making PM resource allocation and workload decisions.

The focus in this chapter is on

1. explaining resource allocation methodologies adopted by industry practitioners,
2. identifying the key project factors that have an impact on predicting staff requirements through analyzing the previous research and discussions with experienced project managers,
3. proposing a data acquisition model to help project managers in collecting projects data in the proper way,
4. collecting historical project data to be used in the development of the forecasting models,
5. data cleaning and transformation to prepare it for further exploration, analysis and datamining, and
6. explaining each factor in the dataset and illustrating the data through graphical representation to show the distribution of factors.

In the next chapter, forecasting model development is explained in detail, including the data preparation process, feature selection and showing the model inputs, and data mining techniques adopted to forecast the owner's PM staff hours required for a given project. Moreover, chapter

four includes models' evaluation and the bootstrapping method which was used to predict the PM hours with a confidence interval.

3.2 Factors Affecting Staff Requirements

There are various factors utilized by the industry practitioners and researchers to forecast the manpower required for a project. One of the objectives of this research is to investigate the possible factors affecting the staff requirements on a project level and to propose a number of factors that can be utilized to forecast the owner's PM staff requirements. In addition to the factors outlined in the literature review chapter, further investigation was conducted to understand the current industry practices. This investigation was performed through discussions with project managers of three different owners to identify the project characteristics to be considered when assigning the number of staff required to a project.

3.2.1 Industry Practices in PM Resource Allocation

Understanding current industry practices is a vital part of this research. The objective is to define the factors considered by project managers to determine the number of PM staff required to manage a project. Discussions with different program managers and project managers working for three owners occurred to understand their resource allocation methodologies. Each organization had a different way of identifying the required number of staff. They adopted both qualitative and quantitative techniques, including heuristic rules based on their experienced managers, regression models from previous projects, and their own framework to rank projects based on multiple project characteristics.

3.2.1.1 Organization #1

This organization acts as an owner that manage infrastructure and buildings projects with total costs from $1 million to $1 billion. The PM staff that manages these projects is categorized into three types: senior project manager, project manager, and project coordinator. This organization has faced some challenges in assigning the required PM staff for a project, including over- and underutilized resources.

The owner adopted a ranking system for projects to determine the number of resources required to manage a project. The project's risk score puts it in a certain category, and the category determines the staff required. There are six project risk categories that are defined: total project cost; project complexity; project delivery method (PDM); number, type, and relationship of stakeholders; impact of poor quality; public or political interest; and geographic location of the project.

Each project feature is divided into multiple categories with assigned scores. The points allocated to each aspect are summed to determine the category that the project falls in. Every project category has a predefined number of PM staff that also includes the number of resources in each resource category. The number of staff and PM resource categories are developed based on the knowledge of experienced senior project managers.

The total project cost represents the level of effort and the difficulty to manage the project. It consists of different classes, each of which is dollar value, e.g. from $0 to $5 million, from $5 to $10 million, etc. The higher the project value, the higher the class and project score.

The complexity of the project is categorized according to the building classification and purpose, and the categories are determined by the type of occupancies in the building and on how the space is used. The objective of this aspect is to measure the effort needed to manage and coordinate the

different trades involved in a project. Although, it is necessary to identify the amount of coordination needed to estimate resources, there is confusion between the project category/type and the project complexity.

Three PDMs are considered in this framework: design-bid-build (DBB), design-build (DB), and construction management (CM). Every PDM has different points due to the involvement required by management to oversite each type of project.

Primary stakeholders are the number of stakeholders involved in the project, and they can be within the organization or another party. It also defines the difficulty that might occur if the stakeholders' goals are not aligned with the owner.

The quality of work distinguishes the effort needed to ensure a high standard of work during execution and upon completion. Moreover, it considers the impact if the required quality is not attained.

The public or political interest aspect accounts for the level of engagement required to communicate and coordinate the information with external parties. These parties will be impacted by the completion of the development of the project and can include consumers, Indigenous people, or political figures.

The geographical location considers the time needed by the project team to travel to the worksite and how far is the site from the PM staff municipality.

The PM resource allocation framework used by this organization provides them with guidance to identify the required resources for an upcoming project. However, the project risk category that is defined by the total points is entirely based on the knowledge and experience of PM staff.

Furthermore, the number of resources specified for each category is deterministic and does not consider the number as distribution which is more realistic. Also, it is not possible to have two projects within the same category have the same number of resources. Finally, depending on a qualitative evaluation without any quantitative analysis provide the project managers with inaccurate results.

3.2.1.2 Organization #2

The organization is an owner who is responsible for projects with values from $10 thousand to $40 million and manages buildings projects. The project management resources are categorized into three types: project managers, project coordinators, and project assistants. They have been facing some difficulties in estimating the required PM staff for a project, and that has affected their employment and outsourcing decisions as the organization's workload exceeded capacity of the available staff.

The organization has been using heuristic rules to estimate the number of PM staff required for upcoming projects. The rule is in the form of minutes per $1000 for any type of project. It was established based on the experience of senior project managers and historical projects (actual project cost and the number of staff on the project). This rule provides them with a good start pointing for the workforce required for a project. However, utilizing this rule for all project types did not provide them with reliable estimates due to over- or underutilization during project execution. Furthermore, the historical data only considered project cost and the number of staff. The number of people who worked on a project is not sufficient without tracking the hours spent by each individual and analyzing the total spent hours on a project. For example, one person can be assigned to two different projects, and quantifying the proportion of time that was spent on both projects is subjective to that individual only.

3.2.1.3 Organization #3

The organization is an owner who manages projects from $10 thousand to $600 million. They oversee different portfolios that include infrastructure and buildings projects. Their project management staff is divided into three categories: program managers, project managers, and project coordinators. One of the main issues facing the organization is to assign and determine the workload capacity of the PM staff.

The organization has various departments that are responsible for the previously mentioned portfolios. Each department estimates the required number of resources based entirely on the knowledge and the experience of the project managers. However, one department developed regression models to estimate the resource requirements using historical projects.

The regression models were developed to allocate the proper number of PM staff for new projects and to determine how busy each individual is. There are three main factors that were considered in the built models: the project phases, the complexity of the project, a calibration factor. The project phases consider the workload required for each project phase, represented in the form of a percentage of the total amount of time that the PM staff need to manage a project. This was calculated based on the hours needed to complete the tasks and the project deliverables of each phase. The percentage of time needed for each phase was implemented for five project classes that were categorized based on the project cost.

The complexity represents the effort that the PM staff need to properly manage a project. The project value was considered as an indicator for the project complexity – when the cost increases the time commitment also increases. As mentioned earlier, the value of the project cannot be the

only indicator for the project complexity; there are other factors that need to be considered such as the technical and the organizational complexities.

The calibration factor specifies whether an employee is currently busy or not busy. For instance, if the percent busy value is 100%, it means that the project manager workload is high and cannot manage additional work.

The hours used for building the models were calculated from multiplying the percentage of the time needed by the total hours spent. Then the estimated hours were plotted against the projects' budget to generate the model. The models were built for projects less than $200 thousand, between $200 thousand and $100 million, and more than $100 million. The regression models were based on only 14 projects, which might provide misleading results.

The models were then validated using another set of historical data, in which the total hours were calculated by dividing the charged costs of the PM staff by the rates. The models have always overestimated the number hours required for a project when comparing them to the actual spent hours. From industrial perspective, the models offer satisfactory results because it is better to have surplus in hours rather than facing shortage of manpower during execution. Although, the models could not be generalized to all types of projects and departments.

The other departments used their own methodologies to estimate the staff requirements. The estimates were based on the experience of the program managers, their familiarity with the projects, and their knowledge about different project types and the amount of effort required. The following aspects were considered in the allocation process:

1. Project total cost. (The higher the project cost, the higher the effort required to manage the project.)

2. Project types. (Each project type requires different amount of time to manage it properly. For example, a bridge reconstruction project does not need the same amount of effort as a resurfacing project.)

3. Public engagement in the planning phase. (Receiving input from the public or other stakeholders consumes more time than the projects that do not require public involvement.)

4. The number of contracts in a project. (Managing multiple contracts in the same project requires coordination by the project manager, which is more effort.)

5. Seasonal work. (The weather impacts certain types of projects, specifically parks and playgrounds projects, which leads to a higher workload in summer and higher fluctuations.)

6. Design work. (In-house design does not require PM staff involvement but outsourcing the design work does.)

7. New or rehabilitation projects. (A rehabilitation project's scope is subject to changes and revisions, impacting estimated hours.)

8. Complexity. (The complexity is based on the dollar value and the project type; the higher the dollar value, the higher the complexity.)

Most of the departments did not track the hours spent on the projects, making it difficult to validate the process of defining the staff requirements. Furthermore, a few of them adopted Excel sheets for the allocation process, without having a proper framework or methodology. The project cost

was associated with the highest weight for classifying the projects, followed by the other factors. A project gets a score based on the effort required and the criticality of the project. Based on the score, the number of resources required is determined.

3.2.2 Industry Practices Discussion

In the PM resource allocation process, various techniques were qualitative or quantitative adopted. However, each of these frameworks have limitations that impacted the forecasting accuracy and led the practitioners to refrain from using their methodologies. These drawbacks can be summarized as follows: solely relying on project managers' knowledge and experience and not analyzing the actual project values compared to the estimated values; the allocation methodologies can not be generalized across departments and project types; and the mistaken belief in representing the complexity as a dollar value or as a project type. However, the project complexity is defined by other factors such as technical and technological complexities, site layout, organization complexity, and unpredictability of the construction site (He et al., 2015; Y. Lu et al., 2015; Wong et al., 2008).

A few project's characteristics that were considered by the project managers to estimate the PM resources are difficult to track and quantify, thus were disregarded in this research. For instance, evaluating the impact of poor quality on a project and determining the level of effort needed to manage that project is extremely subjective. Moreover, the effort required to communicate between multiple stakeholders involved in a project cannot be quantified as it depends on the social relationship between the parties and their flexibility. Since, the projects in the data collected were constructed in the municipal areas of the owners, the projects geographical location was also not considered.

3.2.3 Selection and Elimination of Factors

The factors affecting PM staff requirements were identified by conducting the literature review and determining the factors considered by industry practitioners through understanding their methods in estimating the PM requirements. In this research, the objective is to collect the most important project features from historical data to be able to forecast the PM hours required for a given project.

In the literature review chapter section 2.3, it was mentioned that there were a few forecasting models developed by researchers to estimate the workforce required for a construction project. These models considered multiple factors in the forecasting models such as the actual cost, project type, floor area, etc. and overlooked other factors such as complexity, PDM, and type of work (new or rehabilitation project). Table 1 shows all the factors impacting the project's staff requirements considered by the researchers in developing their models.

Although, other factors such as the project type, complexity and site conditions were disregarded from forecasting model of Wong et al. (2008), they mentioned that these factors should not be overlooked while estimating staff requirements. Additionally, some factors that are mentioned by researchers cannot be generalized to all types of projects or the data was unavailable. For example, the floor area can only be considered when estimating buildings projects. Few of the technical risks, such as subcontracting conflicts and difficulty of acquiring permits, are challenging as they cannot be quantified or estimated in an early stage of the project. Thus, they were overlooked in this study.

Table 1 Factors Impacting Project's Staff Requirements Considered by Researchers

Factor.	Frequency	Reference
Project type	4	(Bell & Brandenburg, 2003; Chen et al., 2008; Persad et al., 1995; Yang & Kim, 2019)
Project cost	4	(Bell & Brandenburg, 2003; Chen et al., 2008; Persad et al., 1995; Wong et al., 2008)
Work type	2	(Othman et al., 2011; Yang & Kim, 2019)
Floor area	2	(Chen et al., 2008; Proverbs et al., 1999)
Project's technical risks (e.g.: site conditions, subcontracting conflicts, permits, etc.)	1	(Yang & Kim, 2019)
Project duration	1	(Chen et al., 2008)

By conducting discussions with multiple project managers who work with the three different owners the following factors were selected to be examined in this study:

1. The project's budget. (The factor represents the estimated dollar value that is expected to be expended to complete a given project.)

2. The project's type. (The factor signifies the type of the construction project that will be built. For example: highway project, medical care facilities project, parkade project, etc.)

3. The project's complexity. (The factor indicates the technical and technological complexity of the construction project, although the practitioners used the dollar value to determine the complexity.)

4. Public engagement. (The factor represents the effort needed to receive input and feedback from the stakeholders.)

5. New or rehabilitation project. (The factor indicates whether the project is new, or renovation work to an existing facility).

Few factors that might have an impact on the total PM hours requirements were not considered because it was difficult to collect this information due to its unavailability of confidentiality. For example, considering the impact of weather on the project's hours was disregarded since it was difficult to collect this type of data and develop a relationship between the weather and the total hours required. Moreover, the number of contracts initiated in a project might be a factor affecting the effort needed by the project managers, but it was difficult to collect any information about the number of contracts due to the unavailability or confidentiality. It is difficult to determine at the beginning of a project the number of contracts that will be signed, which made it unrealistic to involve such a factor.

Two new factors were proposed in this research to signify the site conditions of the project and the effort required by project managers to communicate and manage outsourced design team and multiple contractors. The first one is the field type, and the second one is the PDM. Thus, a total of nine factors was used as an input to the forecasting models in this research. These factors include: project cost, duration, complexity, PDM, category (buildings, industrial and infrastructure), subcategory (e.g. business, recreation, parkade…, etc.), and type (renovation or new); public engagement; and project field type (green or brown). The selected factors were then reviewed by experienced project managers to ensure that these factors made sense. Also, they were asked to provide their feedback if they seen any missing factors that need to be considered.

3.3 Data Acquisition Model for Forecasting PM Staff

Collecting data in the proper manner has become essential in various industries as useful information can be obtained through analyzing the data. Utilizing the information acquired can improve the company's overall performance by increasing the efficiency and reducing the total costs. This will help the company to have a competitive edge in the market. Recently, the construction industry has started to move towards this path to be able to do predictive analyses, productivity improvements, proactive decisions regarding safety, etc. Therefore, obtaining good quality data is crucial for the organization success.

The data collection process involves two stages: determining the information required and the means of collecting it. By understanding the business process of the owners, it was observed that they did not track their projects properly. Therefore, missing information was found when the data was pulled from their tracking systems, and a data acquisition model is proposed to help in collecting the required project information for forecasting the PM staff in a structured way. The model will improve the data collection and the resource allocation processes in an organization.

This framework is proposed to allow the involved organizations to track their project data properly. This outline will support these companies in utilizing the PM resource forecasting model by collecting good quality data. Moreover, it will help them in conducting future analyses that could provide them with useful knowledge to take corrective actions and make proactive decisions. An entity relationship diagram (ERD) was developed to graphically represents the different entities incorporated in the framework and the relationships among these elements. Additionally, the attributes of the entities are included to define the information that needs to be collected for the purpose of this research.

3.3.1 Entity Relationship Diagram (ERD)

The purpose of developing the ERD, shown in Figure 4, is to visualize the relationships among the different entities that need to be tracked and stored in the database. The objective is to track the PM hours spent for each project phase because the effort varies from one phase to another. For instance, the PM staff is more involved during the construction phase than in the planning phase. Though when there is public engagement during the planning phase, more effort is required. Moreover, the number of resources fluctuates during the project, so it is difficult to assume that the hours required for a project is divided equally between the months of the project duration. The proposed framework will support the owners in collecting the required project information for resource forecasting and the most essential attributes of a project to perform other future analyses.

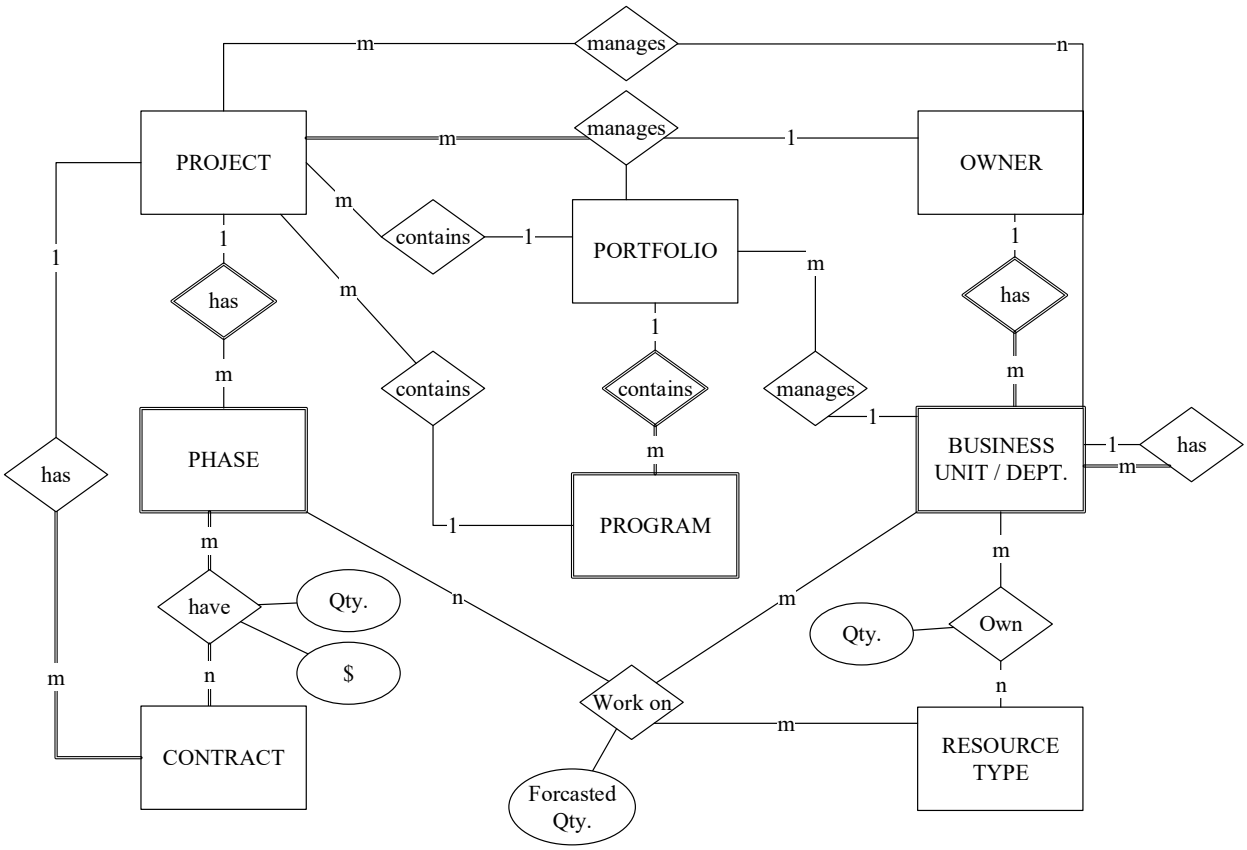


Figure 4 ERD

The model is built for the owners to collect good quality data; hence, an owner entity is included in the ERD that the rest of the entities are built on. Each of the entities mentioned in the ERD has multiple attributes that will be shown in the next section. The owner has multiple business units or divisions which include various departments. The model considers three levels of the organization chart to reach the level at which the projects are managed. Each of department owns different types of resources with several quantities that denotes the available staff. The resources involve project management team, design team, construction supervision team, etc.

Each business unit can manage one or more portfolios to reach the organizational strategic goals and objectives. A portfolio consists of multiple programs that will help achieve organizational objectives through strategic business objectives along the way. A group of related projects form the program; the program's goals cannot be completed without combining these projects. Incorporating project portfolio management is crucial for the success of the organization as it allows the company to assess, prioritize, and choose projects aligned with its overall strategy and assign resources to projects considering the project's priority (Blichfeldt & Eskerod, 2008; Martinsuo & Lehtonen, 2007; Meskendahl, 2010).

3.3.2 Project Attributes Required for Forecasting PM Staff

The project is comprised of several phases such as initiation, planning, design, construction, and warranty. Moreover, there are one or more contracts in a project that provide various services, including consulting, construction, soil investigation, etc. Each phase requires particular types of resources to execute the phase's scope and deliver its requirements. Therefore, the forecasting of staff requirements should be performed on the phase level when the organizations collect the data. Also, there is a relationship between a contract and a phase, as a consultant may be contracted to

perform the design phase and could provide construction support based on the contract, but a contractor is contracted to execute the physical construction.

Based on the collected data, it was impossible to forecast the PM staff required for each phase. Accordingly, the project is the most investigated entity in this research because the forecasting of the staff requirements is reliant on the project characteristics. The project characteristics are formulated in the ERD as the attributes that need to be collected for construction projects in Figure 5. The collected information is not only for the forecasting model; it will provide the owners to carry out proper future analyses which could support them in the decision-making process.

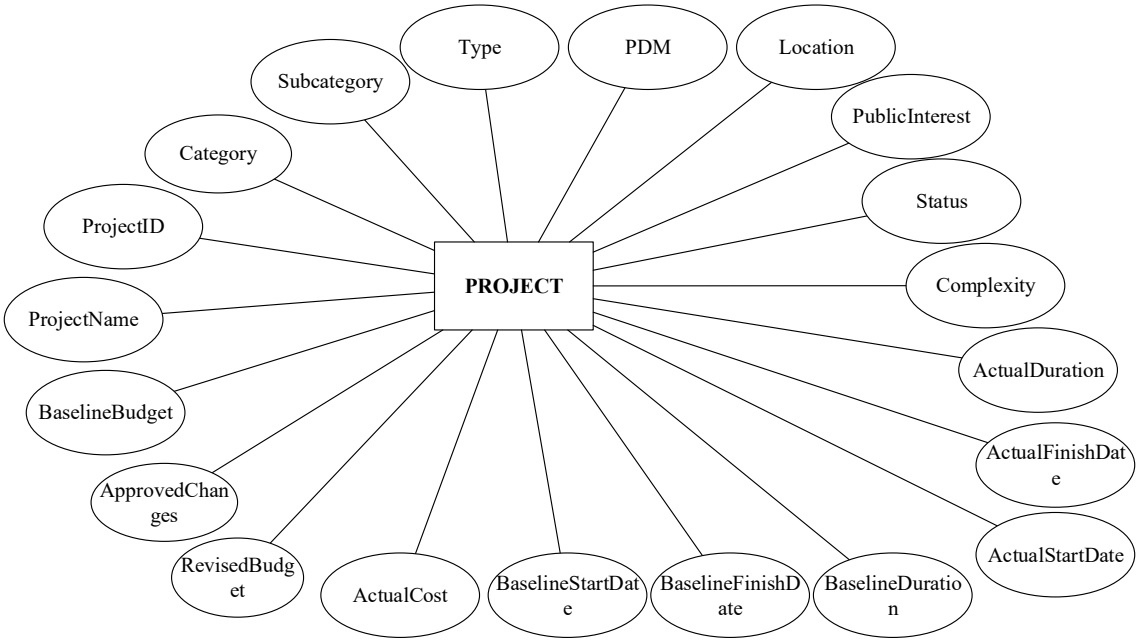


Figure 5 Project Attributes

3.4 Historical Data Collection

The data for training the forecasting model are key for a robust model. In this research, historical project information is collected from three owners that involves different types of buildings and infrastructure projects. The data collected varied from three- to eight-year periods, between 2012 and 2019.

The objective of collecting the data is to utilize it in the forecasting models to estimate PM resource requirements. The historical data play important roles in ensuring the forecasting model works properly. The information that needs to be collected, which impacts the PM staff requirements, were proposed in the literature review and the industry practices. In this research, the purpose is to gather historical data on the project level. There are two methods of data collection implemented in this study: first, collecting objective data in the form of numerical values from enterprise resource planning (ERP) systems, such as SAP ERP and Oracle ERP, and PM Systems, such as e-Builder and Procore; and second collecting the missing information, including the subjective data through discussions with program managers or project managers.

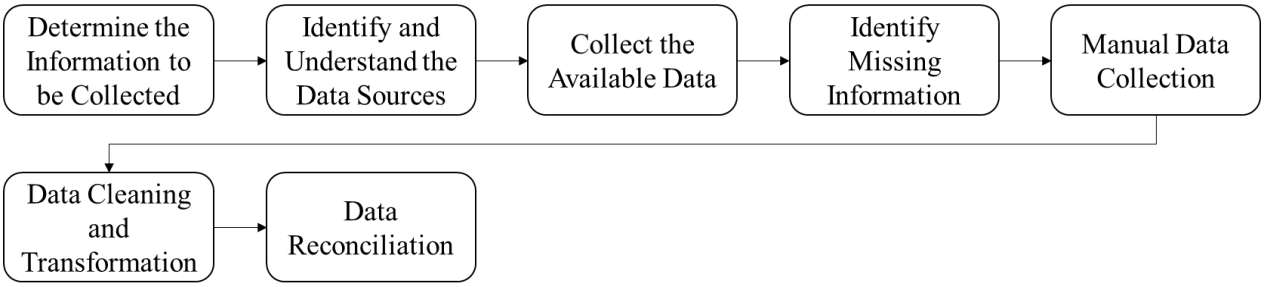


Figure 6 Data Collection Process

3.4.1 Information Necessary for the Forecasting Model

The factors that affect PM staff requirements and need to be collected were determined through conducting the literature review and understanding the industrial practices of three different owners. The following are the crucial factors for a project: project cost, duration, and complexity; PDM; project category (buildings, industrial and infrastructure), subcategory (e.g. business, recreation, parkade…, etc.), and type (renovation or new); public engagement; project field type (green or brown).

In this research, the objective is to collect these project attributes for the historical projects that were completed over the last eight-year period. The PM staff hours spent on these projects were collected to train the forecasting model.

3.4.2 Data Sources

Each owner has its own way of collecting project data that is scattered between the ERP systems, PM systems, and unstructured data (e.g. status reports).

The first owner tracked the projects through status reports in unstructured text documents and spreadsheets. Each manager responsible for multiple projects developed their own tracking sheets. The reports included the project budget, consultants, contractors, actual start dates and finish dates of the project, and the project phase. Another report incorporated the allocated PM staff on each project. The values of PM staff were estimated based on the developed framework. The actuals of the projects were not included in any of the reports and the spent hours on the projects were not tracked.

The second owner had an ERP system and a PM system for resource planning and financial tracking purposes. The project budget, revised budget, and expensed amount for projects were tracked. The expensed amount involved soft costs (any expenses not directly related to the construction cost), and hard costs (expenses that are directly related to the physical construction). Moreover, the hours charged to the projects were tracked using the dollar value spent on the project, including the outsourced resources.

The third owner used an ERP system that tracked the actual project costs, divided into internal and external expenses. The internal expenses involved internal resources charging to the projects and overheads, while the external expenses included the payments to consultants and contractors. The

project budget and duration were in a different system for top management reporting purposes. A PM system was recently introduced to the company to track project schedule, cost, budget, parties involved in the project, type, category, and complexity. The PM system was fairly new, and included a few in-progress or recently started projects. Accordingly, the data were pulled from the other two systems and only included the spent staff hours and the actual cost.

3.4.3 Missing Information and Manual Data Collection

None of the owners tracked factors other than the project's budget, actual cost, planned duration, and staff hours. Moreover, few departments allocated their hours on projects in a proper manner. Attributes such as project type, actual duration, category, subcategory, complexity, public interest/engagement, PDM, and field type were collected manually through discussions with the project managers who worked on these projects. The project managers defined these characteristics by searching their own tracking sheets or based on their memory.

Project complexity was represented on a five-point scale for ease of use. The project managers were asked to provide the ratings based on the project's technical and technological complexities and the unpredictability of the project. However, the determined complexities were subjective, and the gathered ratings are biased towards the owners' experience and portfolio.

The project managers were provided a list of selection for other features such as the project category, project type, field type, PDM, and public engagement. For the project type, they were required to identify if the project was new or a rehabilitation of an existing facility. A renovation project could include doing physical construction and replacing electrical and mechanical equipment. The field type indicated if the project was executed in a brown field (existing facility) or in a green field (new worksite). The PDM were design-bid-build, design build, construction

management, and in-house. In-house category was included as few owners have some trades that carry out small size construction activities. Public engagement was determined by yes or no, as it was difficult to assess the level of effort required to deal with the public in each project. Finally, the project managers selected from buildings, industrial, and infrastructure to determine the project category.

The project subcategory was handled differently as each of the organizations provided their input from the wide classification lists adopted in their entity. Although, their classifications for projects were established by the building purpose or a short description of the scope, it was difficult to collect this type of information differently. There was a wide range of classifications found in the collected data, e.g. medical care facilities, hospitals, fire stations, police stations, sports fields, stadiums, swimming pools, arenas, etc.

3.4.4 Data Cleaning and Transformation

Three datasets were collected, one from each owner. One of the datasets included the charged hours to projects in the form of dollar values. The main challenge for this dataset was to acquire the rates of the PM staff to calculate the hours spent on projects. There were various labor classes who worked on these projects such as PM staff, design staff, and planning staff. Therefore, it was essential to aggregate the spent hours using the labor class, because the objective was to collect the hours of the PM resources. Moreover, a few projects were not charged with hours hence they were removed from the dataset. Additionally, the project duration was not tracked; thus, they were computed through identifying the date of the first and the last transactions of the projects.

The major difficulty in the second dataset was that the owner did not use timesheets for employees working on projects. Therefore, number of staff allocated to projects were used instead of spent

hours. The staff hours were calculated based on the working hours per week excluding statutory holidays. The hours were only computed for the projects that included durations, and the remaining ones were excluded.

The third dataset included actual project costs and internal staff hours spent. The data were pulled by the profile number which consisted of multiple projects and the data were exported into 70 Excel sheets. The ID of each line item consisted of 2 IDs, the project ID and the cost breakdown structure ID; thus, the challenge was to aggregate the data by the project ID and include only PM staff hours. First the PM hours were pulled from the sheets. Then, the IDs were manually separated to include the project ID only. Finally, the mapping between the projects list associated with the actual cost and the hours was performed.

Another obstacle in the collected data was the presence of various project subcategories for each project category. Every company had its own method of classifying projects, and each department or division managed certain classes. For example, one department was responsible for building health facilities, while another in the same organization was for governmental facilities. Another owner had departments that manage facility projects, medical care facilities, and transportation projects. Some of these projects were classified differently in these organizations; however, they had comparable scopes and purposes. Therefore, it was critical to propose a generic classes for the project subcategories to be able to utilize this attribute in the forecasting model because it is a significant factor that impacts the total staff hours. The 2018 International Building Code (IBC) was utilized to develop a classification for each project category based on the building's or the structure's occupancy and designated space use. Also, the suggested subcategories/classifications represent the levels of hazard and risk to building's tenants and the primary function of the building. The 2018 IBC classified the buildings/structures into 10 categories:

1. Assembly (theaters, restaurant and casinos, recreation and worship, arenas and stadiums).
2. Business (airport traffic control towers, post secondary schools, banks, barber and beauty shops, car wash, etc.).
3. Educational (schools and daycare facilities).
4. Factory and Industrial (low-hazard and moderate-hazard factories).
5. High Hazard.
6. Institutional (supervised care facilities like alcohol and drug centers or assisted living facilities; medical care facilities like foster care facility, detoxification facilities, or hospitals; high security facilities like correctional centers and detention centers; and adult and child daycare facilities).
7. Mercantile (department stores, drug stores, markets, etc.).
8. Residential (residential buildings, hotels, motels, etc.).
9. Storage (storage facilities and low-hazard storage facilities).
10. Utility and miscellaneous (agricultural buildings, aircraft hangars, barns, carports, grain silos, etc.).

The subcategory of each project was selected by matching the project with the description of each classification provided by the 2018 IBC. A few project subcategories were not included in the 2018 IBC such as infrastructure projects. Furthermore, factory, industrial and high-hazard buildings were modified to be categorized under industrial projects, unlike the 2018 IBC.

To develop the target data prior to importing it into the forecasting model, the collected data from the different systems was imported into an MS Access database. In the MS Access database, the data is well organized, structured, and easily manipulated and managed.

3.4.5 Data Reconciliation

Mistakes can be made while migrating the data from the various sources into one database. Accordingly, ensuring that the data was transferred and mapped properly was crucial to avoid dropping records, missing and incorrect values, and duplicated records.

The numerical values found in the database were verified against reports, which were pulled out from the ERP and PM systems and status reports. The manually collected data were considered accurate because the values were directly filled in the database. Additionally, the actual cost and the spent PM staff hours were plotted against one another to confirm the correlation between them based on the literature review and the industry practices. This helped in validating that the mapping was done correctly, and the columns' values were not misplaced.

3.4.6 Data Description and Exploration

Nine key project factors impacting staff requirements were identified. Then, these factors were used as inputs into the forecasting models. The following section will briefly discuss the nine factors and how the data is distributed for the different values of each project attribute.

3.4.6.1 Total Project Cost

The total project cost represents the dollar value spent to complete the project. The higher the total cost of the project, the bigger the project size is. The dataset involved projects of total cost ranging between $10 thousand to $1.4 billion. However, 95% of the projects were below $50M. Figure 7 shows the actual cost of projects collected from the owners. Projects with total cost greater than

$50M were disregarded from Figure 7 since they were only 20 projects. Another point to consider is that nearly 420 projects in the dataset had total cost less than $3M as shown in Figure 7.

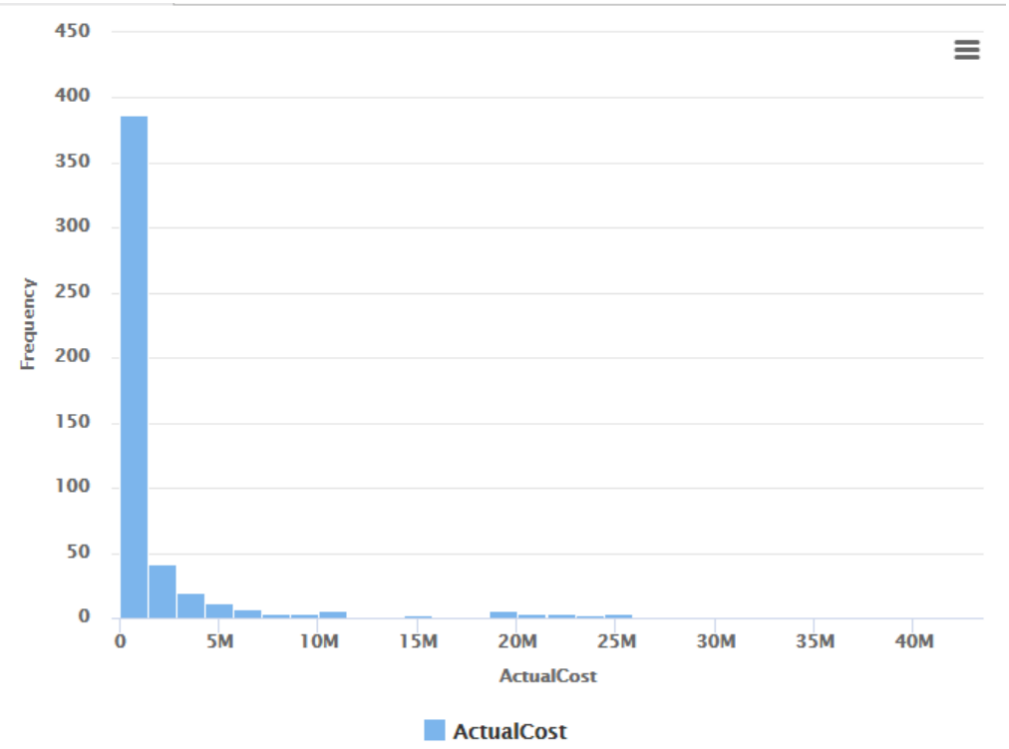


Figure 7 Distribution of Projects' Total Cost in the Dataset

3.4.6.2 Project Duration

The project duration is in months and it represents the time needed to finish the project. The objective of this factor is to quantify the impact of project duration on the total PM hours required to finish the project even if the project's size was small. Figure 8 shows the distribution of projects duration in the dataset.

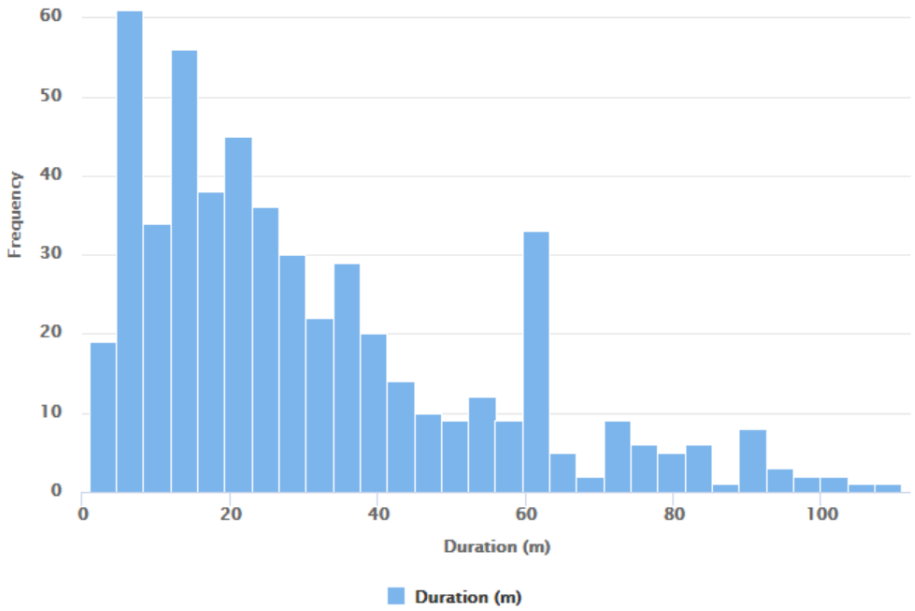


Figure 8 Distribution of Projects' Duration in the Dataset

3.4.6.3 Project Complexity

The complexity of the project is represented by a five-point scale to define the project's technical and technological complexities and the unpredictability. The objective of this aspect is to measure the effort needed to manage and coordinate projects with different level of complexities. Figure 9 shows the distribute of projects' complexities in the dataset.

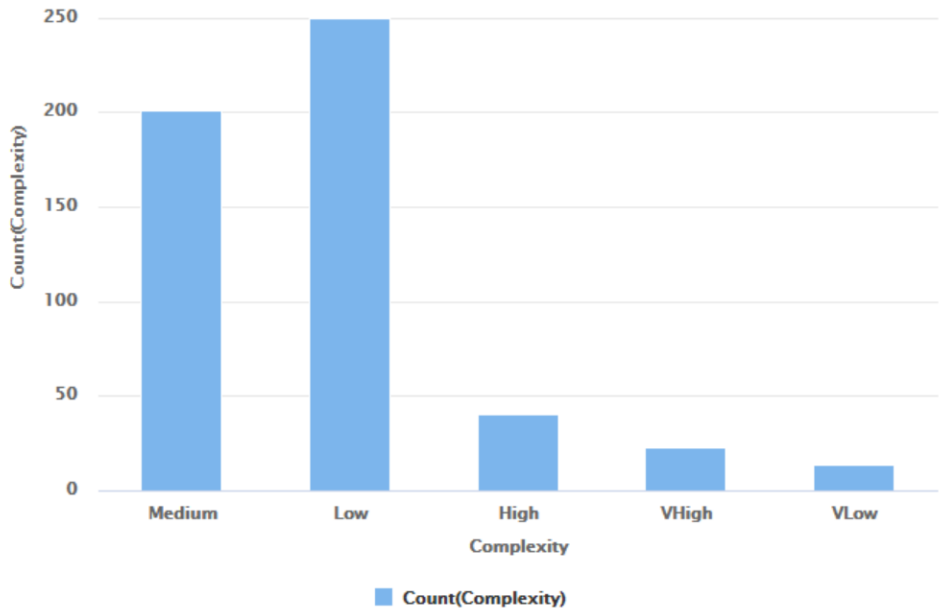


Figure 9 Distribution of Projects' Complexity in the Dataset

3.4.6.4 PDM

There were four PDMs considered in this research: DBB, DB, CM and In-House. (In-House signifies if the design and the construction work is done by the owner's workforce.) The PDM was proposed in this research because it can represent the effort needed by the owner's PM team to manage and coordinate between the designer and the contractor. Since each PDM requires different level of involvement by the owner in the design and the construction. Figure 10 shows the number of projects that fall under each PDM in the dataset.

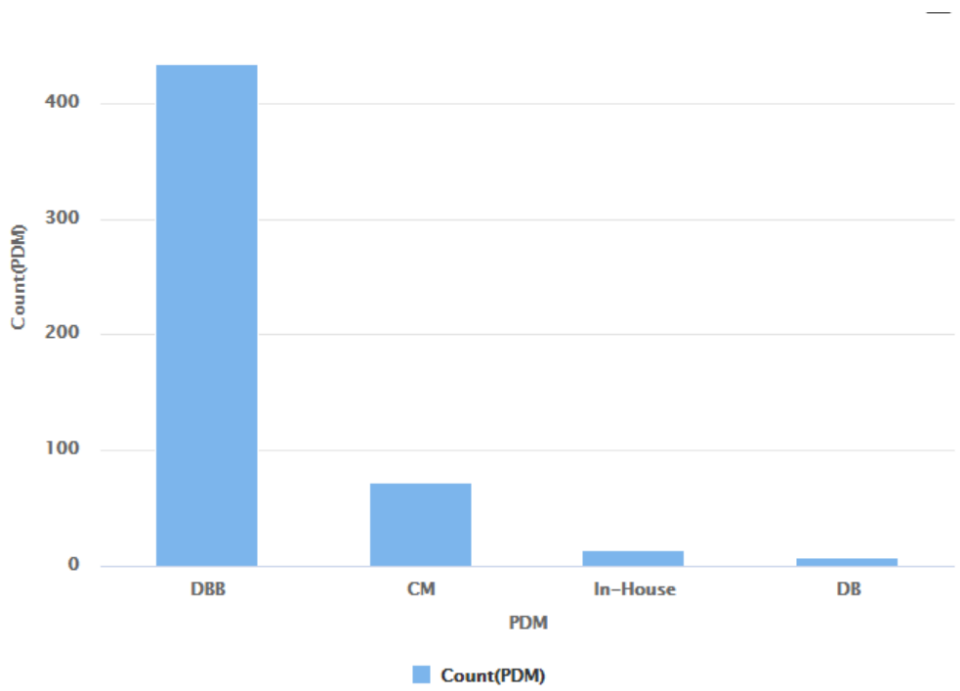


Figure 10 Distribution of PDM in the Dataset

3.4.6.5 Project Category

Project category represent the type of the project that will be built such as buildings, infrastructure, or industrial project. Buildings projects include medical care facilities, residential, business, etc.; infrastructure projects involve roads, bridges, parks and playgrounds, etc.; and industrial projects include oil and gas, power plants, factories, etc. Figure 11 shows the number of projects within each category and as illustrated most of the projects collected were buildings and infrastructure projects.

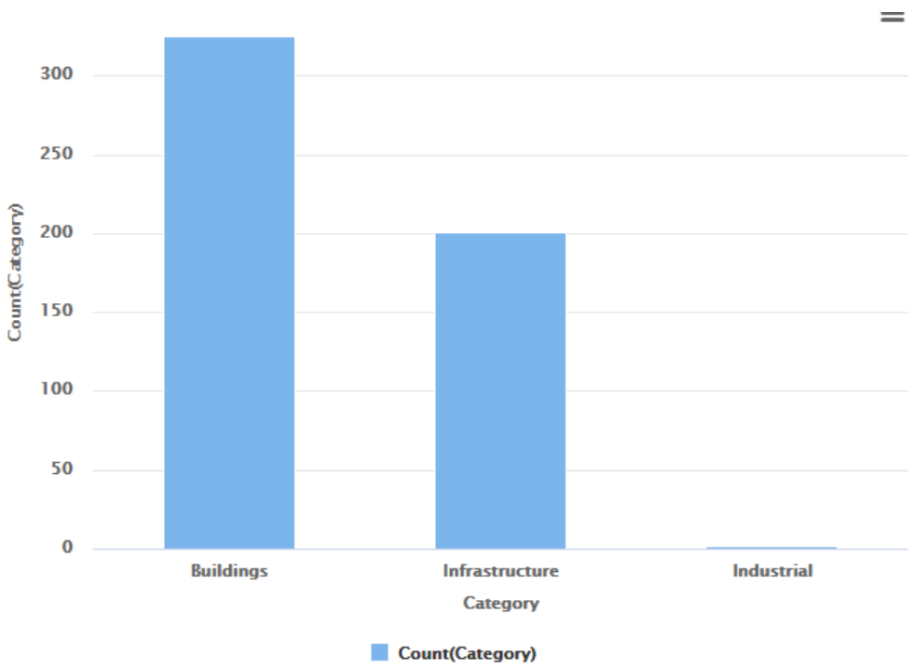


Figure 11 Distribution of Projects' Categories in the Dataset

3.4.6.6 Project Subcategory

The project subcategory defines the type of project within each project category. For example, business buildings, residential buildings and medical care facilities fall under the building's category. Another example is parks and playgrounds, bridges and neighborhoods fall under the infrastructure subcategory. Figure 12 illustrates the subcategories involved in this research and stating the number of projects falling under each subcategory.

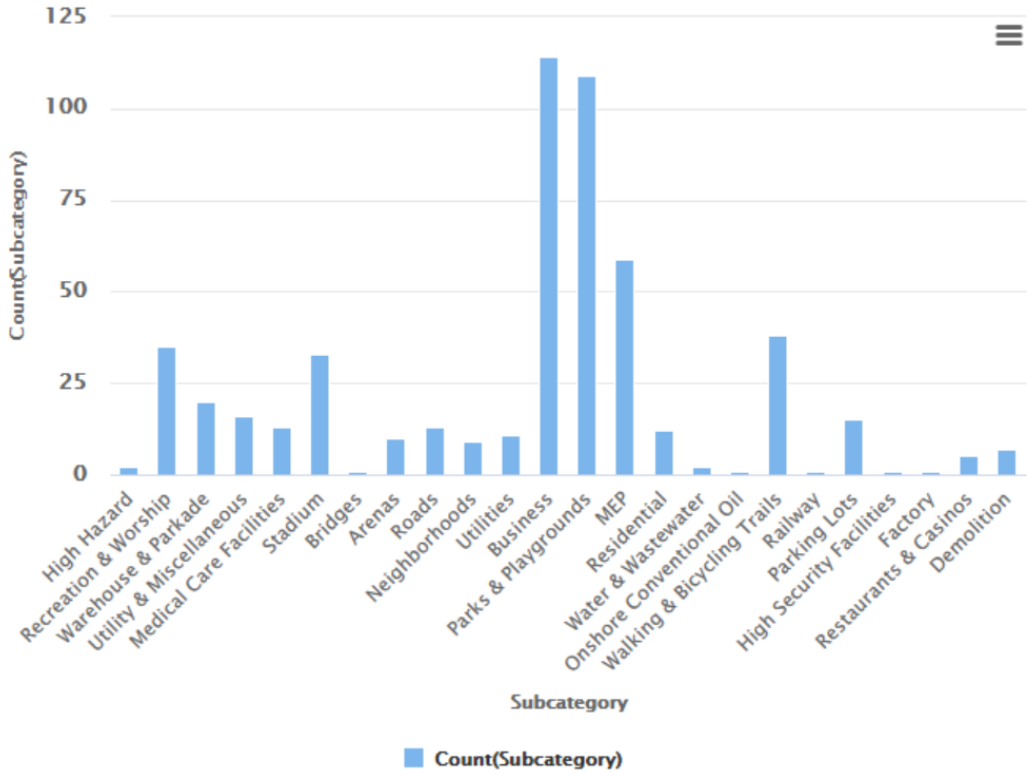


Figure 12 Distribution of Projects' Subcategories in the Dataset

3.4.6.7 Project Type

The project type represents the type of work that will be performed in the project whether it is a new project or rehabilitation project. Each project type has its own complexity in terms of the technical and managerial challenges that will face the project manager managing that project. Figure 13 shows the number of projects under each project type in the dataset.

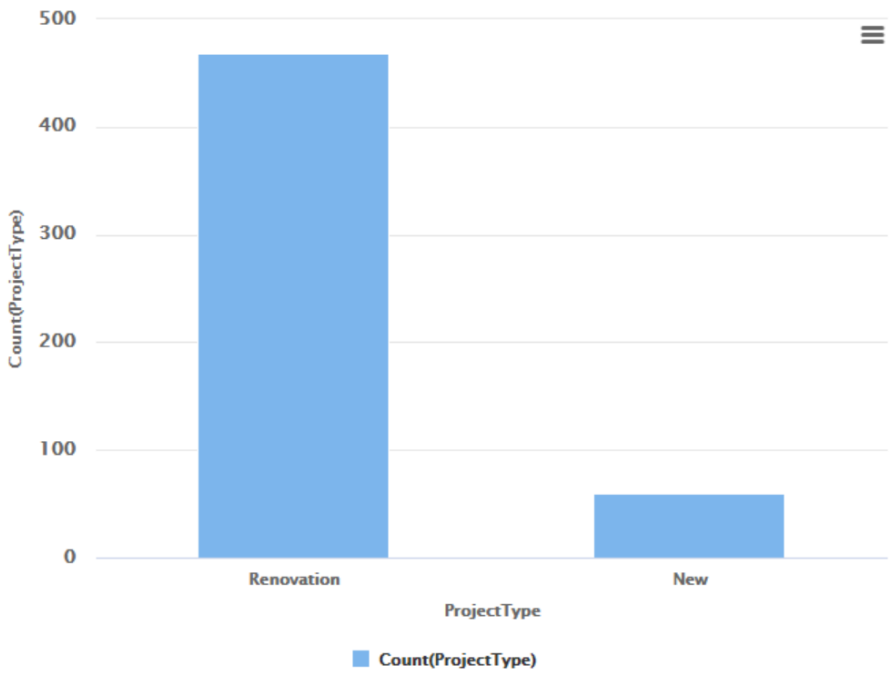


Figure 13 Distribution of Project Type in the Dataset

3.4.6.8 Field Type

The field type defines the site conditions of the project. There are two values used in this research: green field that represents an empty worksite free of any existing facilities and brown field that indicates that the worksite contains existing facilities and it is challenging to work in. The data collected from the owners involved both green fields and brown fields. However, all renovation projects were in a brown field, and all new projects were in a green field. Thus, one of these factors was used as an input to the forecasting model since it was redundant. Figure 14 shows the number of projects that were built in brown and green fields in the dataset.

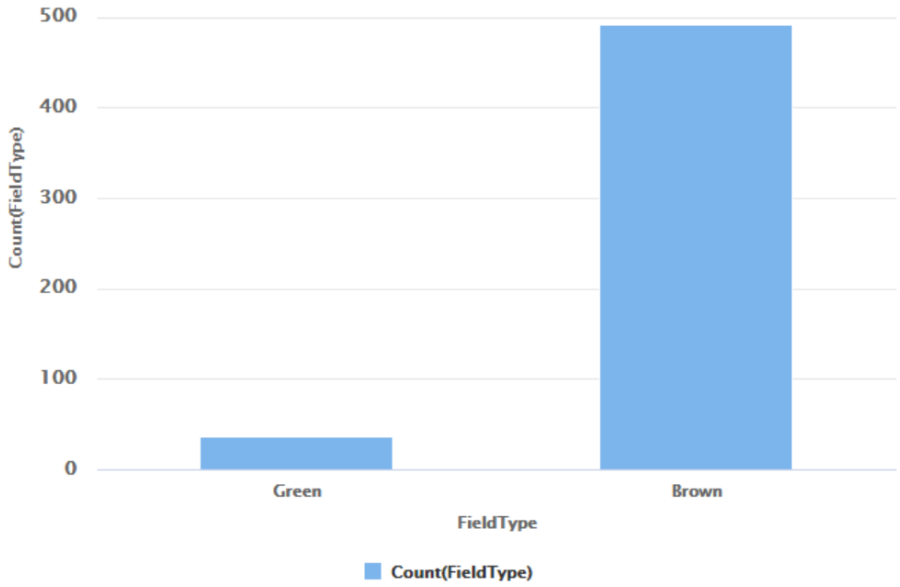


Figure 14 Distribution of Field Type in the Dataset

3.4.6.9 Public Engagement

The public or political interest aspect accounts for the level of engagement required to communicate and coordinate the information with external parties. At this stage of the research, it was difficult to quantify the level of the public engagement since it was historical data that was tracked. Therefore, true and false values were used to indicate the public engagement occurred during project execution. Figure 15 shows the number of project that involved public engagement and the projects that did not involve any public engagement.

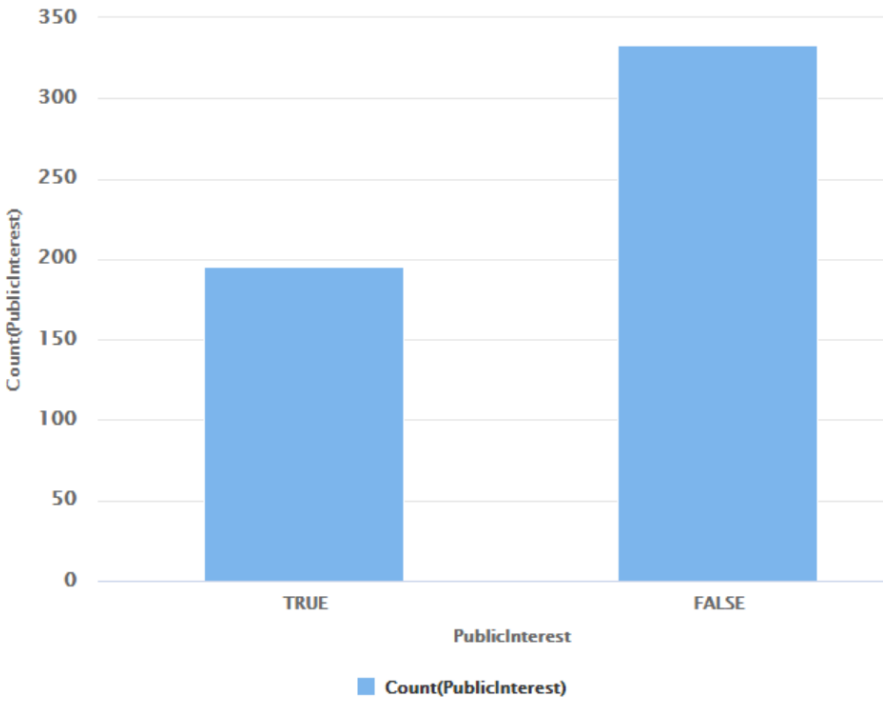


Figure 15 Distribution of Public Engagement in the Dataset

3.4.7 Limitations in the Dataset

There are a few limitations that were encountered during the data collection process. The limitations are present due to the improper tracking systems utilized by the owners and ignoring the tracking of a few critical project attributes such as duration and spent hours. Therefore, some assumptions were made to have a target dataset to forward in this research. These limitations include the following:

1. Calculating the actual project duration of an owner's dataset by subtracting the first transaction date from the last transaction date. This is only an estimate of the actual project duration and cannot be considered as the actual duration of the project.

2. Using the budget and planned dates due to the unavailability of the actuals in an owner's dataset. They were utilized to increase the size of the dataset and improve the forecasting model.

3. Computing the staff hours of a project using estimated staff numbers, working hours per week excluding statutory holidays, and planned duration in one of the datasets.

4. Improper tracking of hours spent on projects. For example, some departments do not charge hours to projects and others do. Another example is when a project is over budget, the staff charge their hours to an under-budget project instead.

5. Calculating the actual staff hours, in one of the datasets, by dividing the actual cost for the PM resources by the rates of these resources.

6. The manually collected data relied on the memory of the project managers.

7. Project complexity ratings are biased towards the owners' experience and portfolio.

Chapter 4: A Forecasting Model for PM Staff Requirements

4.1 Introduction

The objective of this research is to forecast the number of PM staff required for a construction project utilizing historical data, to identify significant project factors affecting total PM staff requirements using feature selection, and finally, to support project managers in making decisions regarding the workload and project assignments of their current resources.

In the previous chapter, nine factors impacting PM resource allocation were identified. In this chapter a methodology for forecasting the PM hours using these factors as input was proposed. The concept behind this methodology was to group projects that have similar characteristics together so the forecasting model can perform better and minimize the prediction error.

The research objective was achieved by applying data mining techniques to develop an accurate forecasting model. Figure 16 shows the process adopted to create the forecasting model. First, unsupervised learning (k-means algorithm) was used to find the patterns in the data and group projects that are within the same cost range and spent PM hours, and detect the anomalies found within each group of projects in the dataset. Additionally, feature selection approaches were implemented, such as the filter method, embedded method, and wrapper method, to determine which project factors need to be considered to minimize the forecasting error. Multiple linear regression, random forest, KNN and ANN algorithms were evaluated to determine which algorithm performed better in forecasting the PM hours required for an upcoming project. Finally, a solution was proposed to overcome the single point estimate of the machine learning algorithms by applying the bootstrap method.

Figure 16 The Process of Developing the Forecasting Models

In this chapter, data preparation, outlier detection, and cluster analysis are discussed. The feature selection approaches are illustrated, and their results are demonstrated. The developed forecasting model is explained, and the comparison between the different regression models is conducted. The forecasting model testing and validation are covered as well. Finally, bootstrapping is used to provide range estimation for the PM staff hours required for an upcoming project.

4.2 Data Preparation

Many of the machine learning algorithms require some data preparation to obtain useful knowledge. Since the algorithm depends on the input data, the data need to be formatted in specific ways, and the algorithm could provide misleading or inaccurate results if the utilized data is invalid. This process involves removing or replacing the null values found in the dataset because of the improper tracking and outliers that are present due to the lack of quality control in tracking and storing the data. Then, the dataset needs to be scaled to avoid the model from assigning more weight to a factor compared to the others because of its greater value. Finally, handling the categorical attributes is another essential part of machine learning as some of the algorithms require numerical input such as regression, support vector machine, and neural networks.

The input data need to be properly formatted to avoid misleading or inaccurate results. Several steps were taken in this process to have legitimate input data for the model. This process involved removing projects with missing values, eliminating outliers, handling categorical project attributes, and scaling the input data.

4.2.1 Null Values

Some of the projects' collected data did not include the project duration or the start and end dates, and the total PM staff hours of other projects were null. This occurred due to lack of allocation of hours to projects on timesheets or the absence of timesheets to track the spent hours on projects. Accordingly, these projects were removed from the input data as it was unfeasible to collect this information in any other way. This reduced the size of the input data, since more than 200 projects did not include this information.

4.2.2 Outliers Detection and Cluster Analysis

At first, a forecasting model was developed using an ANN in which the model was trained on the whole dataset after data cleaning and null value removal. The model did not perform well and provided high error predictions that were worse than using averages. Thus, further data exploration was done to determine impacts to the model's performance. It was observed that the spent PM staff hours in several projects were less than 100 hours. However, Bell and Brandenburg (2003) stated that the minimum hours required by a project manager to manage a project and execute the usual tasks is 100 labor hours (LHR). Furthermore, project managers' opinions regarding the necessary LHRs were considered. It was concluded that the projects with labor hours less than 100 should be removed as the spent hours were not properly allocated to these projects. For example, on some projects that were over-budget, the PM staff allocated their hours towards other under-budget projects; in other cases, the PM staff did not track their hours on a regular basis. Table 2 shows the mean absolute percentage error (MAPE) and mean absolute error (MAE) of the ANN model prior to and after removing projects that are less than 100 LHRs. It should be noted that the MAPE and number of projects are significantly reduced after removing the outliers; however, the error is distinctly high.

Table 2 ANN Model Error Results

No. of Projects	Adjustment	MAPE	MAE
528	–	± 683 %	± 1043 LHRs
361	Exclude <100 LHRs	± 257 %	± 1545 LHRs

The improper allocation of hours has a large impact on the forecasting model's performance due to the presence of misleading data. Therefore, another outlier detection phase was implemented to improve the forecasting model and remove invalid project information. It was difficult to identify outliers using the entire dataset due to variance in projects' actual cost and PM total staff hours. Accordingly, cluster analysis was applied to group projects based on the total cost and PM hours. These two factors were considered in the clustering as they are highly correlated to each other. Higher cost projects correspond to higher PM hours required to manage and execute these projects. Also, a PM hours ratio (the number of PM hours spent on a project per $10,000 of project cost) was proposed to facilitate the process of identifying outliers along with the spent PM hours on projects and actual cost. Spent hours alone could be deceiving. For example, a low-cost project might have higher spent hours compared to similar projects, but it is not easily detected as an outlier because the hours are within the range of a higher cost projects.

The PM hours ratio was used to plot a normal a distribution showing the mean and the variance of the ratio in the entire dataset. As shown in Figure 17, the ratio varies between 0.01 and 250 hours per $10,000, which is a wide range and demonstrates the presence of anomalies in the data. One of the owners used a similar ratio to estimate the number of resources required; however, after looking into the historical data, their multiplier was not near to any of the actuals.

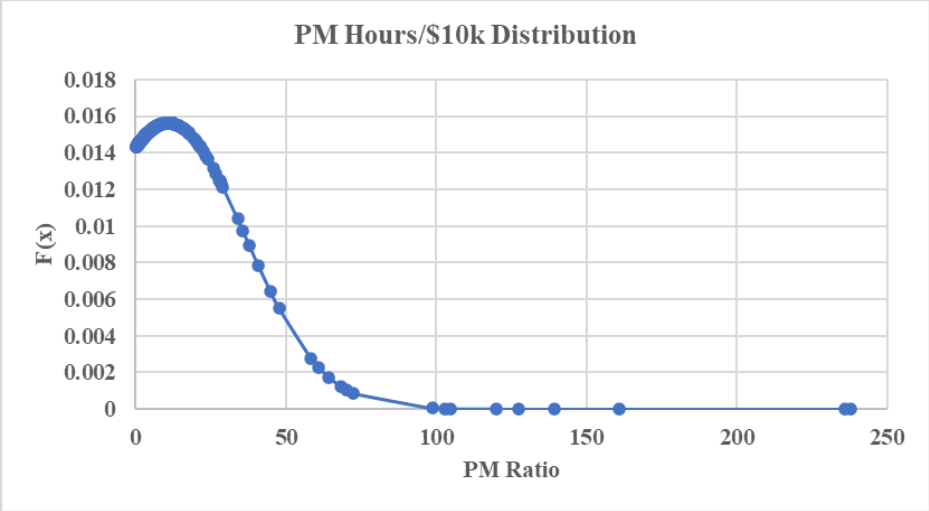


Figure 17 PDF for PM Hours Ratio

Since the ratio distribution did not form a perfect normal distribution and had a long tail, a box plot was used to graphically represent the data through quartiles and identify the outliers visually. The data points below or above the whiskers are outliers to the dataset as shown in Figure 18. About 40 projects with more than 17.5 hours per $10K were removed from the dataset, as they fell above the whisker. One example was a project of $600K with 3000 spent hours allocated to, an outlier compared to the projects that were in the same cost range.

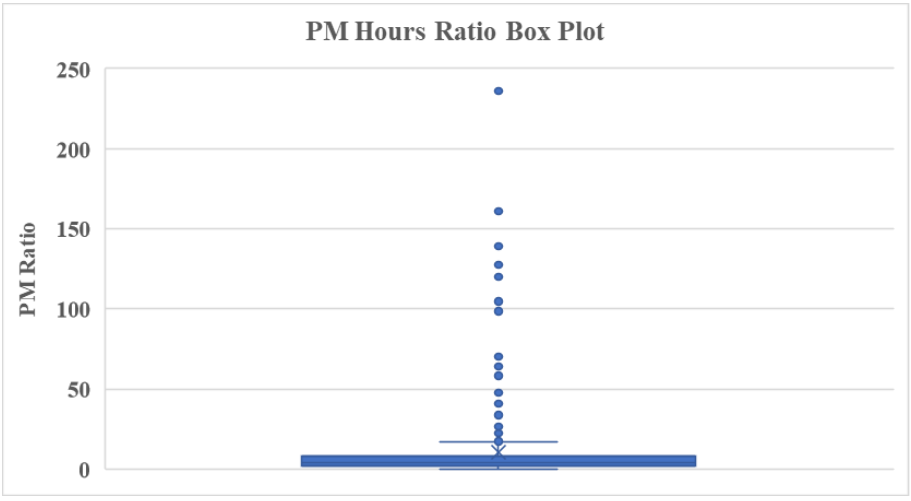


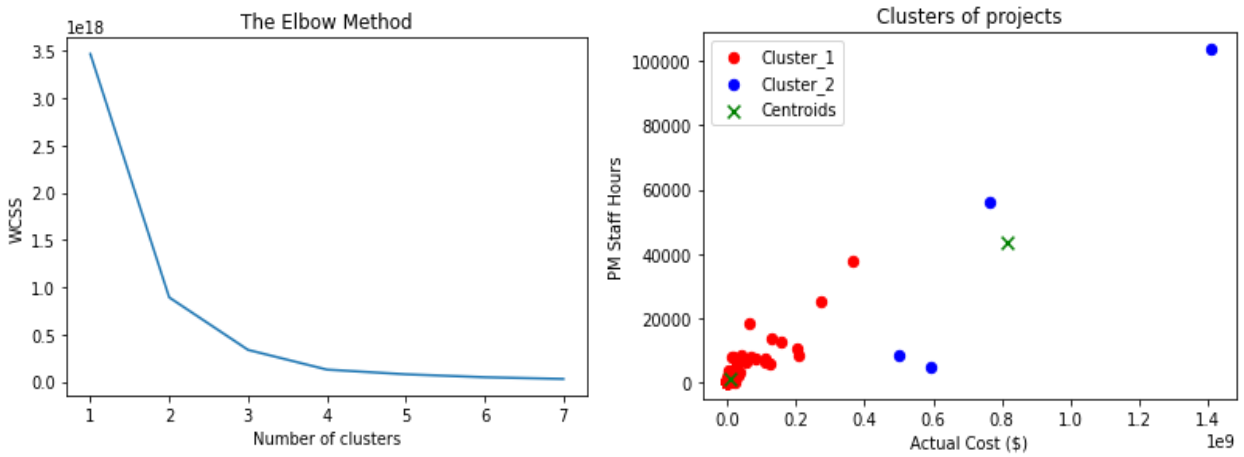
Figure 18 Box Plot for PM Hours Ratio

The next step was to detect the anomalies within each cost range. A cluster analysis was applied to divide projects into different classes based on total cost and total PM hours. K-means algorithm was used on the dataset to categorize the historical projects into a specified number *k* of clusters. K-means is an unsupervised machine learning algorithm that measures the distance between the cluster's centroid and the data objects to determine the similarities. The algorithm creates clusters based on the *k* specified by the user, which might not be the correct number of clusters. Therefore, the elbow method and silhouette analysis were used to validate and select the right number of clusters.

The idea behind the elbow method is to apply k-means on the dataset for a range of *k* values and calculate the within-cluster sum of square (WSS) for each value of *k*. Then, the WSS is plotted against the *k* values used. When the number of clusters increases the WSS decreases towards zero until the number of clusters are equal to the data points. The objective is to pick a small value of *k* that also has a small WSS, and the points with the higher *k* values after the elbow provide a slight improvement to the WSS. Thus, the location of the elbow in the plotted chart represents the best value of *k*.

The elbow method might be ambiguous because the plotted curve is smooth, and the elbow cannot be identified. This occurred a few times with the current dataset. Accordingly, the average silhouette approach was utilized to determine the optimal number of clusters, as well. This approach evaluates the quality of clustering by measuring how similar each point in one cluster is to points in the neighboring cluster. The silhouette values lie between –1 and 1; the closer the value to 1 the farther the points are from the neighboring cluster and the closer to each other. The average silhouette is calculated for a range of *k* values, and the optimal number of *k* is the one with the highest score.

The clustering analysis was done several times on the dataset until the elbow method's curve was smooth, and the silhouette score decreased towards zero. Using a range of *k* values between 1 and 7, the WSS was obtained and plotted to obtain the elbow location. Two clusters were formed after trial number 1, grouping projects more than $500 million and less than $500 million as shown in Figure 19 B. The silhouette score was not used in the first trial as the elbow location was easily identified on the plot Figure 19 A.



(A) (B)

Figure 19 Clustering Trial #1

Based on the first trial of clustering, the projects that are above $500 million were excluded from the dataset prior to the second trial. Focusing on projects that are less than $500 million, another clustering analysis was performed, and both the elbow method and silhouette score were used to obtain the optimal number of clusters. The silhouette score was introduced from the second trial as it was difficult to locate the elbow on the curve. The confusion was between *k* value of 2 or 3 as shown in Figure 20 A; therefore, the average silhouette score was calculated for both to determine which one has a value closer to one. The average silhouette score for both values of k

were nearly close but k equals 2 was slightly higher of value equals to 0.95 (Figure 20 C & D). The two clusters that were formed are projects of actual cost less than $150 million and projects between $150 and $500 million (Figure 20 B).

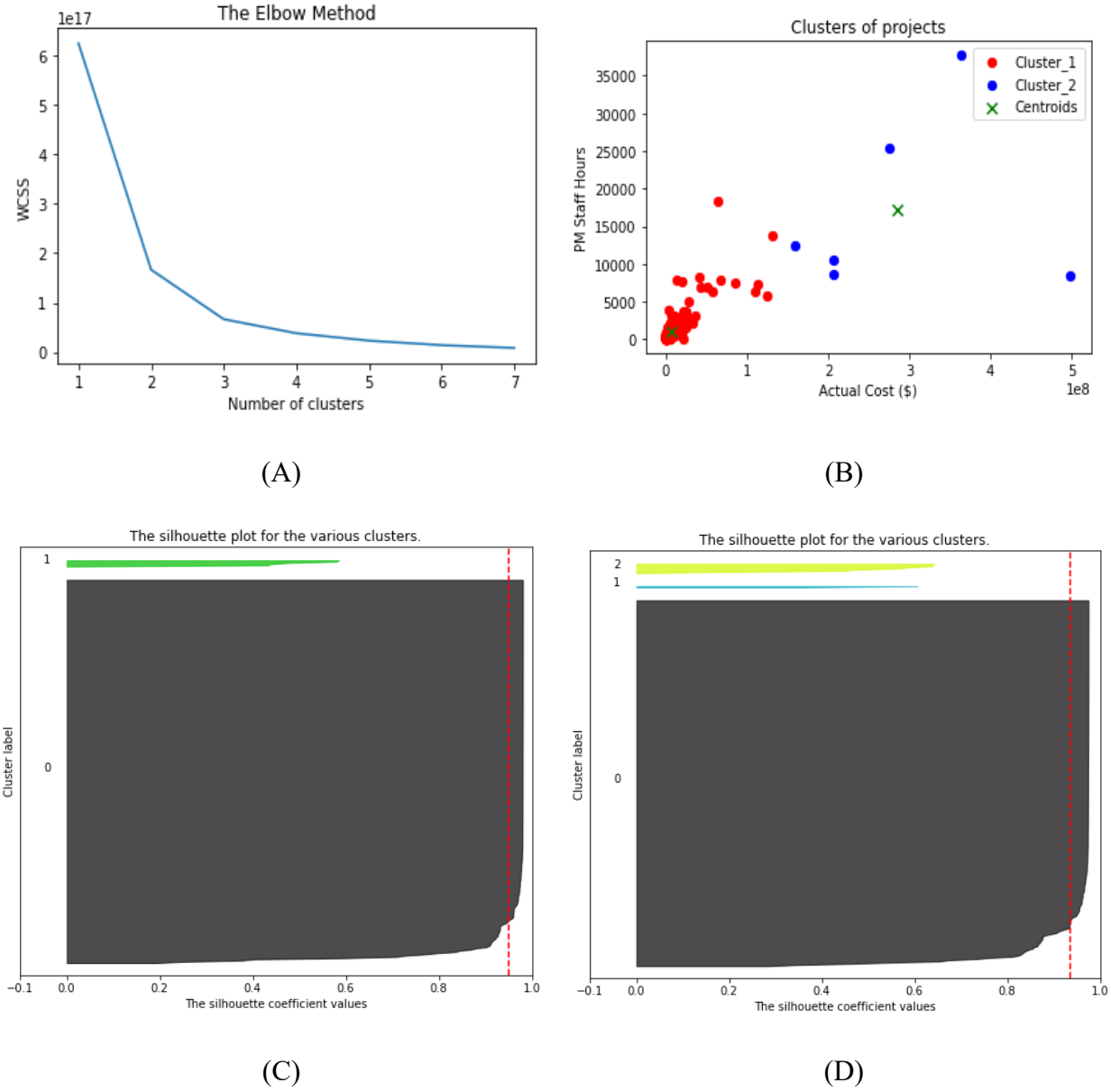


Figure 20 Clustering Analysis Trial #2

Another trial of cluster analysis was applied after removing projects with costs between $150 million and $500 million from the dataset. The optimal number of clusters was equal to 2 as the

elbow was located on the curve at *k* value equals 2 (Figure 21 A) and had an average silhouette score of 0.91(Figure 21 C & D). The projects were divided into two groups: those less than $50 million and projects between $50 million and $150 million (Figure 21 B).

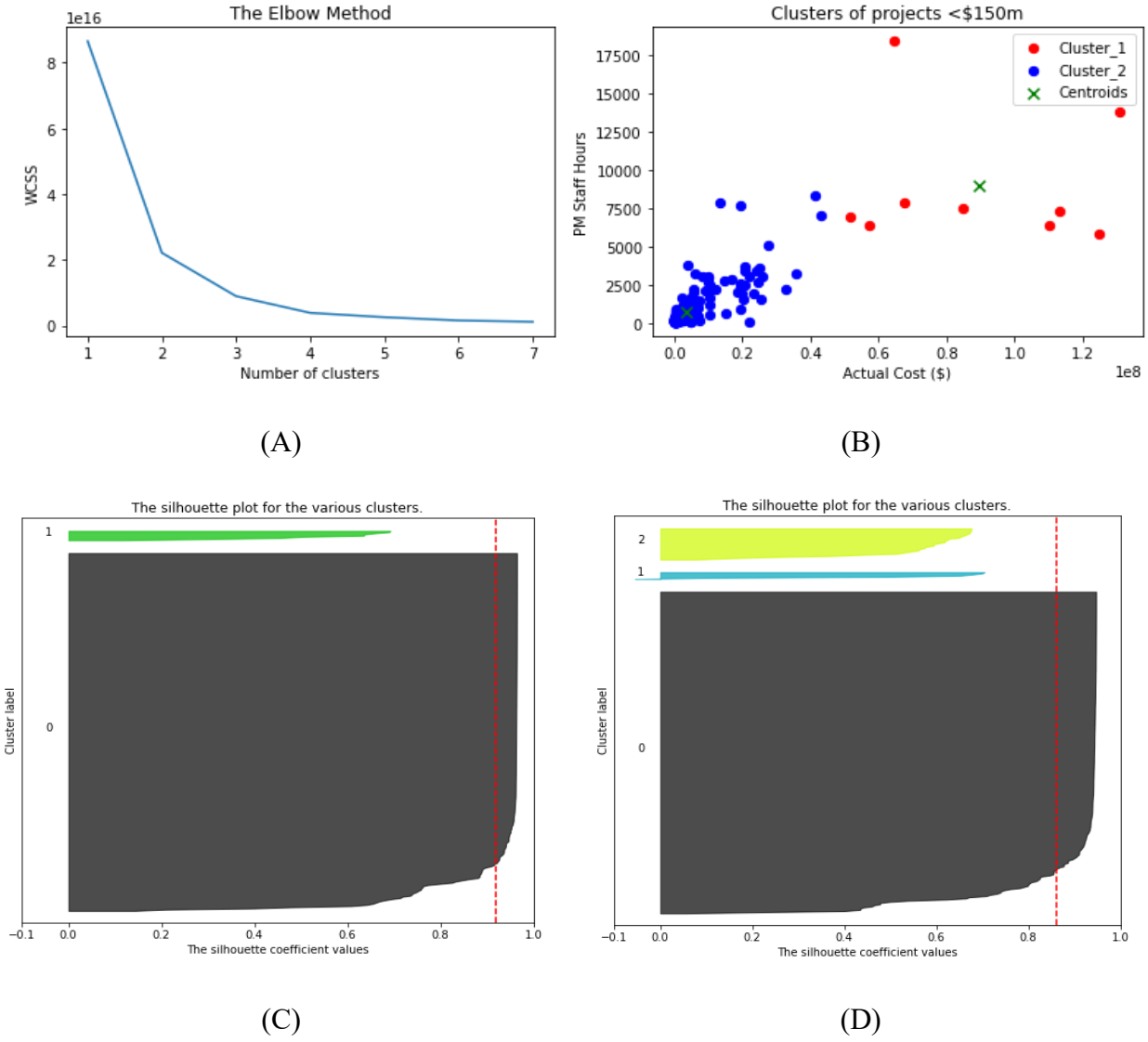


Figure 21 Clustering Analysis Trial #3

Cluster analysis trial #4 involved projects that had a total cost less than $50 million. The best *k* value was equal to 2 based on the elbow method (Figure 22 A) and the average silhouette score. The average silhouette score was 0.86 at *k* value equals to 2 compared 0.79 for *k* equals 3 (Figure

22 C & D). The projects were divided into projects that are less than $12.5 million and between $12.5 million and $50 million (Figure 22 B).

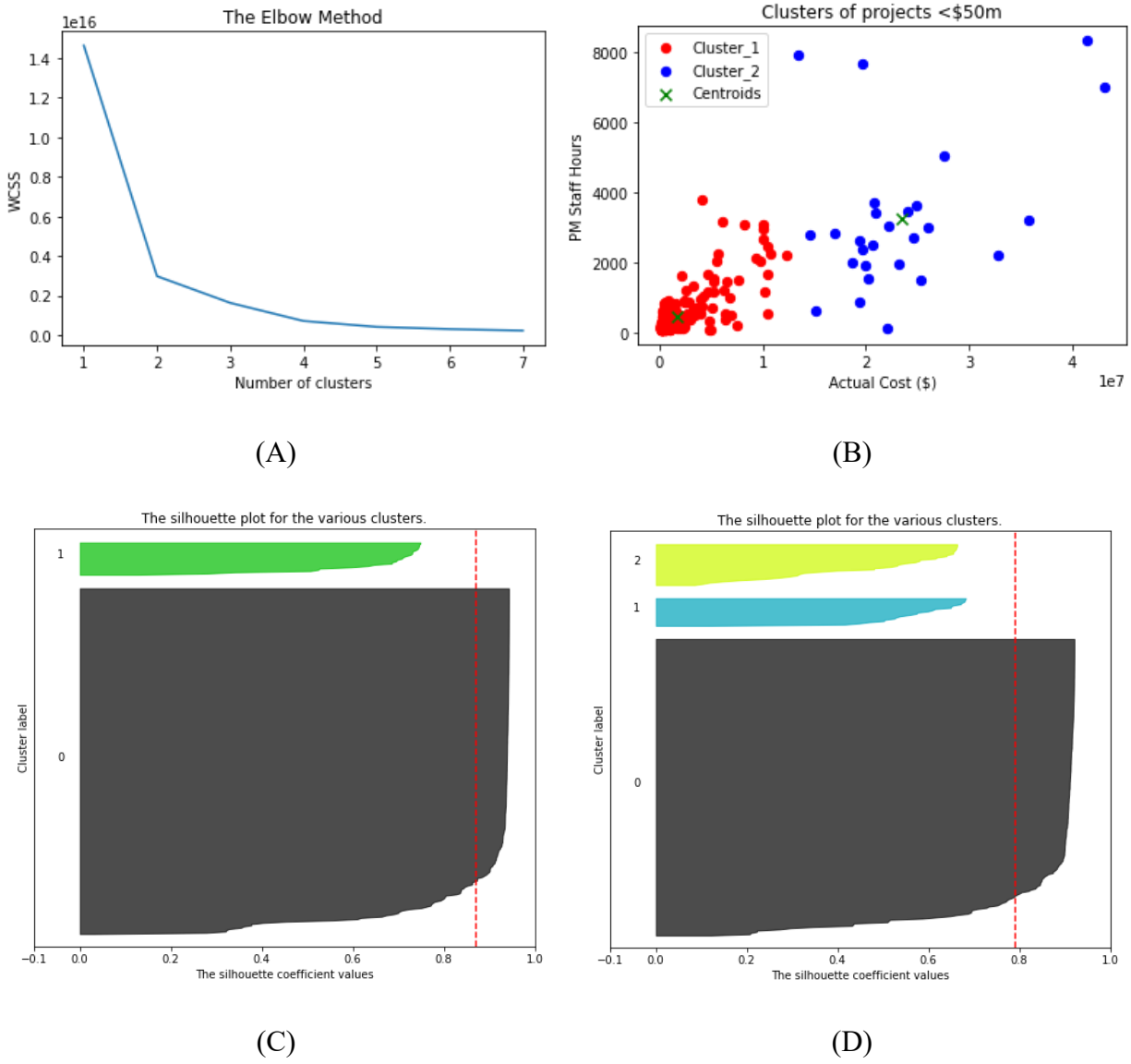


Figure 22 Cluster Analysis #4

Another trial was employed for projects that cost less than $12.5 million. The best *k* value was equal to 2, having the best average silhouette score (Figure 23 C & D), and the elbow located at

that point on the curve (Figure 23 A). Projects were divided into two groups: projects of actual cost below $4 million and between $4 million and $12.5 million (Figure 23 B).

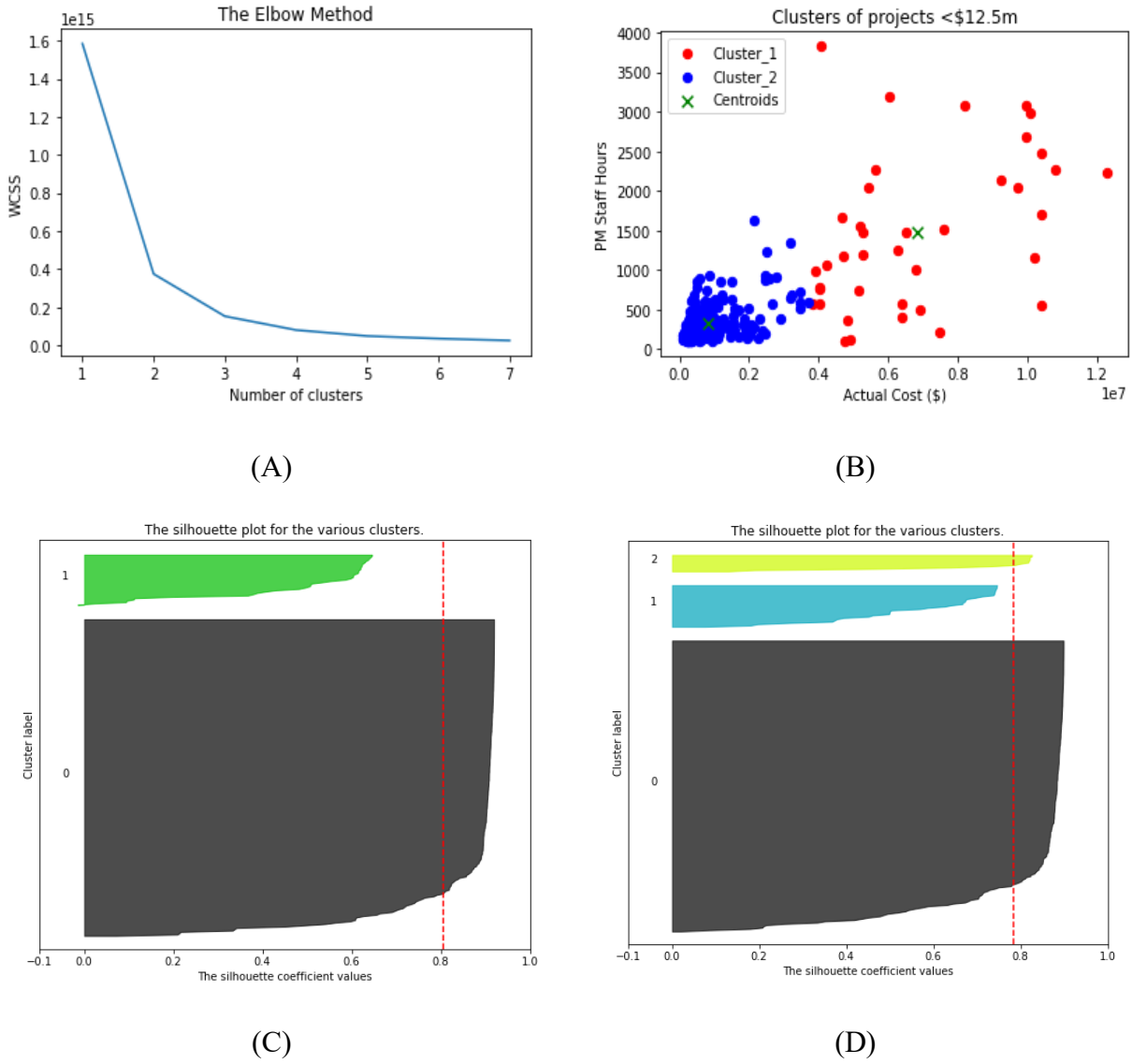


Figure 23 Cluster Analysis #5

Finally, the last cluster analysis was decided when the average silhouette score was declining towards 0, on projects of actual cost that are below $4 million (Figure 24 C & D). The number of clusters was equal to 2 at which the average silhouette score was equal to 0.73 and the elbow at *k*

equals 2 (Figure 24 A). The clusters included projects of actual cost below $1.4 million and between $1.4 million and $4 million (Figure 24 B).

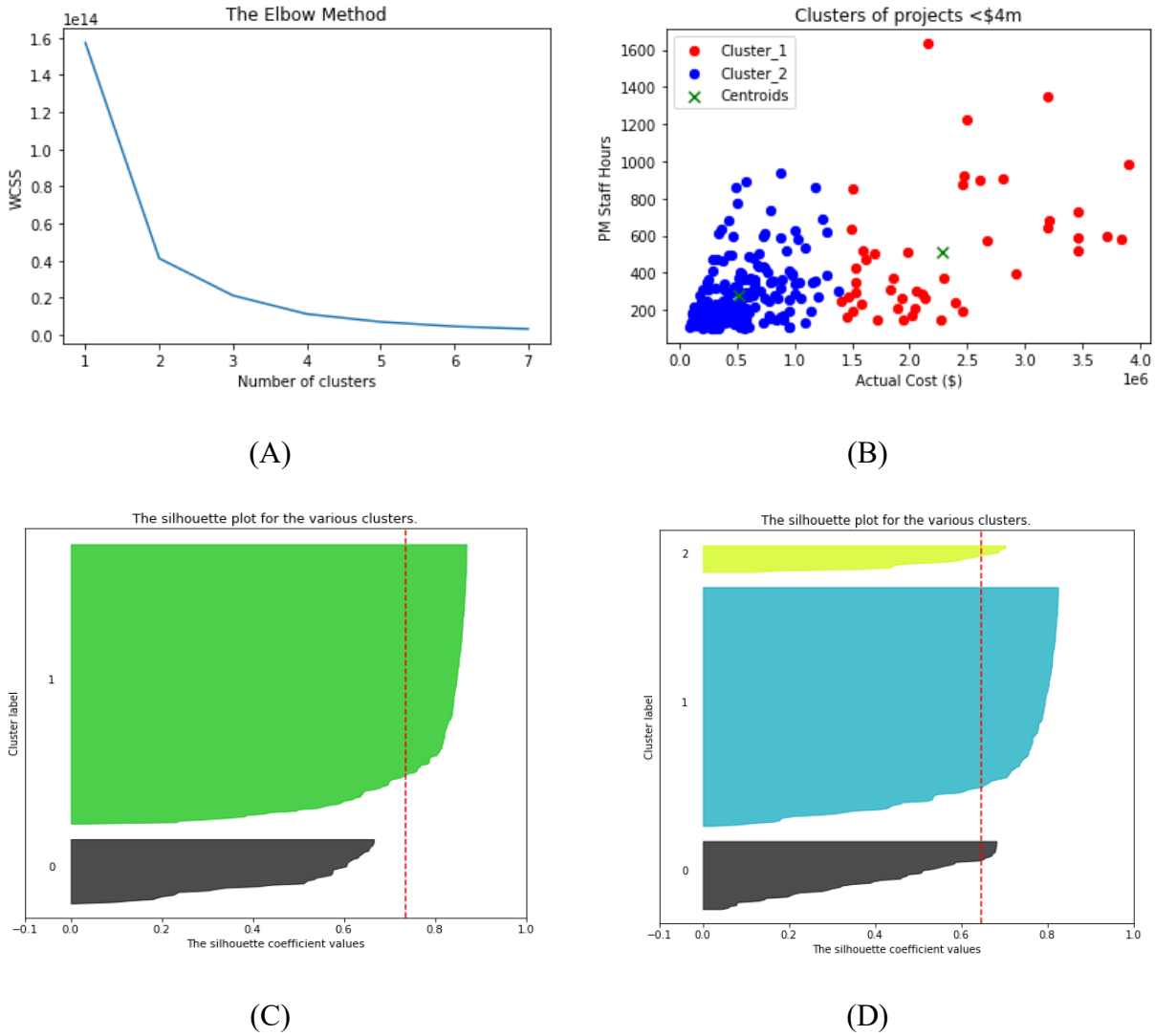


Figure 24 Cluster Analysis #6

After completing the cluster analysis, outlier detection was done on each of the cluster groups to remove projects that would impact the forecasting model. Box plots, scatter plots, and standard deviations were used to detect anomalies found in the dataset. The box plot for the PM hours ratio was used, along with the scatter plot of the actual PM hours against the actual cost. Both methods were valuable in spotting outliers that had higher PM hours ratio and total PM staff hours only.

For several projects that had lower total PM staff hours, those two methods were not able to distinguish them. Thus, the standard deviation method was introduced, which assumed that 70% of the projects in each cluster lie within one standard deviation of the PM hours ratio mean.

The PM hours ratio mean value is higher for the clusters at the lower cost range and decreases when the projects actual cost increases. For example, the PM ratio mean for the cluster of projects less than $1.4 million was 6.7 with a standard deviation of 4.1, where as the $1.4–$4 million cluster and $4–$12.5 million cluster ratios were 2.2 and 2.3 with standard deviations of 1.4 and 1.7, respectively (Figure 25). Within each cluster, this method was used to eliminate projects that lie outside the standard deviation to develop a reliable forecasting model. Table 3 shows the formed clusters including the number of projects and the lower and upper bounds of PM ratio.

Table 3 Projects' Clusters

Cluster	***No. of Projects***	***PM Ratio Lower Bound***	***PM Ratio Upper Bound***
< $1.4M	142	2.6	10.7
$1.4M–$4M	43	0.8	3.6
$4M–$12.5M	31	0.6	3.9
$12.5M–$50M	25	0.4	2.6
$50M–$150M	9	-	-
$150M–$500M	6	-	-
> $500M	4	-	-

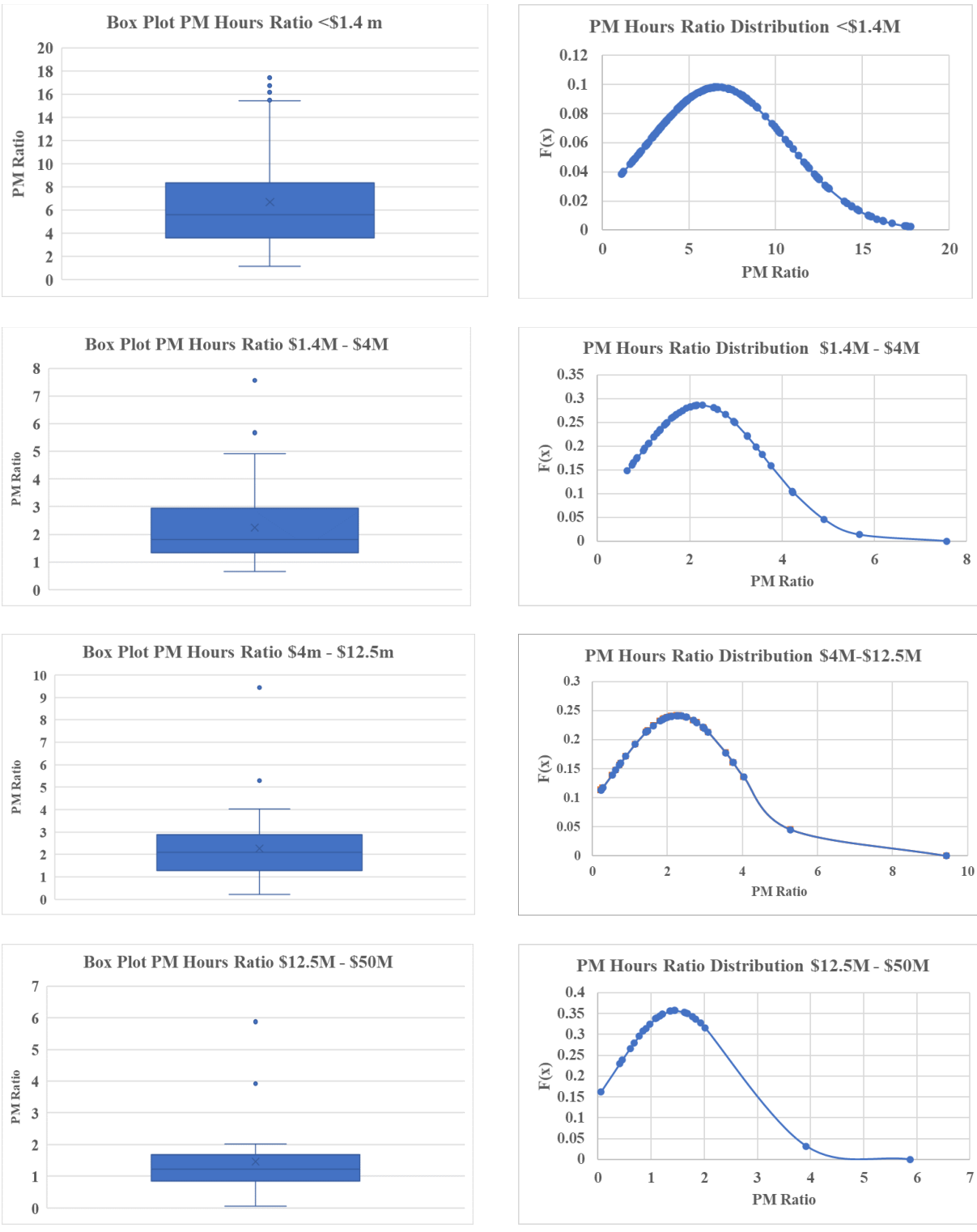


Figure 25 Outlier Detection Process

4.2.3 Categorical Features

Transforming categorical variables is an essential process of data preparation. The collected data included two types of categorical data: nominal attributes, such as project category, subcategory, PDM, type, and public interest; and ordinal attributes, such as the project complexity. Nominal attributes consist of discrete categorical values that are labeled without order. On the other hand, ordinal attributes include categorical values that have an order between them where the distance between these values have an impact on the forecasting model.

Dummy coding was used to transform nominal attributes to numerical values. This method transforms the categorical features into dummy variables. The dummy variable takes a value of 0 or 1. For example, a new column will be created for projects with PDM equal to design build and will have a value of 1; the remaining PDM values will be 0. All categories within each categorical attribute are assigned a dummy value. This method has an advantage over transforming the values into unique integers, as the numerical values cannot be misinterpreted from the prediction or the feature selection algorithms. However, the main disadvantage is that more columns are added to the dataset.

The project complexity involved ordinal attributes with values between very low and very high. Therefore, assigning the proper numerical values to the 5-rating scale was crucial for a good forecasting model. A value from 1 to 5 was assigned to the different levels of projects' complexity, so the algorithms could understand projects with higher complexity should have more weight allocated to them.

4.2.4 Features Scaling

After transforming categorical features into numerical values, all the values in the dataset became numerical. The features varied in magnitude as each feature has a different unit. The machine learning algorithms do not consider the units; they rely only the magnitude of the values. Thus, all project features were standardized to allow the machine learning algorithm to work properly. This is achieved by scaling each of project features to have a mean of zero and a standard deviation of one. So, all the values in each feature were set to a value between –1 and 1.

4.3 Feature Selection and Model Inputs

Feature selection, or attribute selection, is the process of selecting the relevant features when developing a forecasting model. In this research, feature selection eliminated the irrelevant variables to reduce the number of project features and improve the model's performance. In other cases, feature selection can reduce the computational time of the forecasting model, such as when dealing with large dimensional datasets. Moreover, several feature selection methods were implemented to evaluate the output of each method used in the forecasting models and assessing its performance.

Different feature selection methods were used to identify which project attributes have a significant impact on the PM staff hour requirements. The output from each method was evaluated by only considering the selected features as an input for the forecasting models and calculating the errors. First, the filter method was used, which identifies correlations among the project's features and the PM staff hours. Features were ranked based on the correlation coefficient. Then, the wrapper method was used, which involves different techniques such as forward selection, backward elimination, and recursive feature elimination. This method utilizes a machine learning algorithm to identify the best features trying different combinations. Forward selection is a type of stepwise

regression that starts with no variables and then adds the variables one at a time until reaching the best model performance with the fewest variables. Backward elimination is the opposite of the forward selection in which the model starts with all the possible variables. Recursive feature elimination works by fitting the model and eliminating the weakest features using model coefficients until reaching the specified number of features. Finally, random forest feature selection, which is a type of the embedded method, was used for feature selection. This method trains a machine learning algorithm and then derives the features that impact the prediction.

Prior to the feature selection process and transforming categorical data into numerical values, the dataset included nine project attributes. As previously mentioned, the dummy coding process increased the number of columns in the dataset. Thus, before implementing the feature selection process on the dataset, the number of project attributes increased to 26. Feature selection reduces the number of attributes by eliminating irrelevant attributes and improves the performance of the forecasting models.

Feature selection was applied to each of the clusters that were developed using the k-means algorithm. Table 4 shows the project attributes that have a significant impact on the required PM hours in each cluster. The filter method failed to identify the best features compared to the wrapper and embedded methods because the statistical correlation in the filter method requires large datasets. The selected features were evaluated using different machine learning algorithms to measure their performance and determine which features provide better results. For the defined problem, the wrapper method identified the best features which provided the least errors in training and testing the prediction models. Forward selection, backward elimination, and recursive feature elimination were used in the wrapper method and the features shown in the table are the ones that had the best results.

Table 4 Feature Selection Results (Model Inputs)

Cluster	Filter Method	Wrapper Method	Embedded Method
< $1.4M	Actual cost	Actual cost	Actual cost
		Complexity	Duration
		PDM (DBB)	Complexity
		Subcategory (Walking & Bicycling Trails, Parking Lots)	Public interest
$1.4M–$4M	Actual cost	Actual cost	Actual cost
		Complexity	Duration
		Duration	Complexity
		Subcategory (Roads)	Public interest
$4M–$12.5M	Actual cost	Actual cost	Actual cost
		Subcategory (High Security Facilities, Demolition)	Duration
			Complexity
			Public interest
$12.5M–$50M	Public interest	Public interest	Actual cost
		Duration	Duration
		PDM (CM)	Project type
		Subcategory (Business, Roads)	Complexity

Table 4 shows the most significant project attributes resulted from the feature selection process. The wrapper method and the embedded method outperformed the filter method in all aspects. The filter method requires a bigger dataset and perform much better with linear data. Each cluster involved different group factors and these factors were the input to the forecasting model. For

example, the inputs for the model developed for <$1.4M cluster were the actual cost, complexity, PDM and subcategory (the features selected from performing the wrapper method).

4.4 Forecasting Models

Various prediction models were used to forecast the number of PM hours required for an upcoming project. Linear regression, ANN, KNN, and random forest were used for the defined problem. Linear regression was used due to its simplicity and ability to predict with minimal error. KNN was used because of its lack of sophistication and its effectiveness in making accurate predictions. Random forest and ANN are two of the most powerful machine learning algorithms that are widely used in different industries. Also, they are very formidable in dealing with complex problems and nonlinear data.

The models were trained using the collected historical data to discover important trends in the data and relationships between the target variable (PM staff hours) and project features. The first model was developed to forecast the required PM hours for a project, using all projects in the dataset. However, the model had more than 80% relative error. This huge error was present due to the large ranges of project costs and spent PM hours found in the dataset. Big variations in the projects' actual cost led to worse model accuracy. Consequently, cluster analysis was carried out to group projects that have similar actual cost and PM staff. After the clusters were formed, feature selection was done to determine the significant project attributes in each project's cost range. Finally, forecasting models were developed for each cost range using the output of the feature selection, and their performances were compared to the initial forecasting that included all projects.

After identifying the most significant project attributes from the feature selection process, these attributes were the input to the models. They were used to train and test the forecasting models to

predict the target variable, which was the PM hours spent on the project. Multiple linear regression, ANN, KNN, and random forest models were developed, and their performance was compared to each other to determine which machine learning algorithm fit the defined problem.

4.4.1 Linear Regression Models

A multiple linear regression model was developed for each cluster, and their performances were evaluated. The process of building these linear regression models did not require any configurations in the model's hyperparameters, compared to building a neural network model. The hyperparameters are the model's parameters whose values are used to control the models learning process. The models require numerical input variables, which were used to predict the PM staff hours by assigning weights to every variable. The results of the models are included in section 4.5 Models Evaluation.

4.4.2 ANN Models

An MLP neural network with BP algorithm was used in this research. ANN models were developed for the classified clusters as well. However, the model development process was challenging as the model's hyperparameters configuration impacts the model's performance. The process of identifying the number of neurons in the hidden layers was time consuming and is one of major difficulties in creating an ANN structure.

Vujičić et al. (2016) compared different existing methods that determine the number of hidden neurons and concluded that the result of each method is dependent on the dataset and its size. The following three methods were used to estimate the number of neurons in hidden layers, and the results were used to build the ANN.

The number of inputs varied for each model due to the feature selection process performed on each cluster. Since the number of inputs was different, the network topology was different as the number of hidden neurons is a function of the number of variables in the input layer as shown in Table 5. The equations were used to develop these models and determine the number of neurons and hidden layers. Table 6 shows the ANN structures developed for each project range. The structure was developed by evaluating the performance of the model and deciding which ANN structure provides the least error.

Table 5 Methods for Determining Number of Hidden Neurons

No.	Equation	Method
1	$N_h = \frac{(4N_i^2 + 3)}{N_i^2 - 8}$	Sheela and Deepa method
2	$N_h = \frac{\sqrt{1 + 8N_i} - 1}{2}$	Li, Chow and Yu method
3	$N_h = N_i - 1$	Tamura and Tateishi method

- N_h: number of neurons in the hidden layer
- N_i: Number of neurons in the input layer

Table 6 ANN Models Structure

Cluster	Input Variables	Hidden Layers	Number of Neurons
< $1.4M	5	1	4
$1.4M–$4M	4	2	5, 5
$4M–$12.5M	3	2	4, 4
$12.5M–$50M	5	2	6, 6

As mentioned in the literature review chapter, the connections between the input variables, hidden neurons, and output variables are composed of weights. These weights can vary between negative and positive values. The importance of the input variable can be demonstrated by values of the weights between that node and the hidden neurons. To identify the importance of each input variable within the ANN model, the average of the absolute weights connecting an input node with the neuron in the hidden layer was calculated. Figure 26 shows the ANN topology for the <$1.4M cluster.

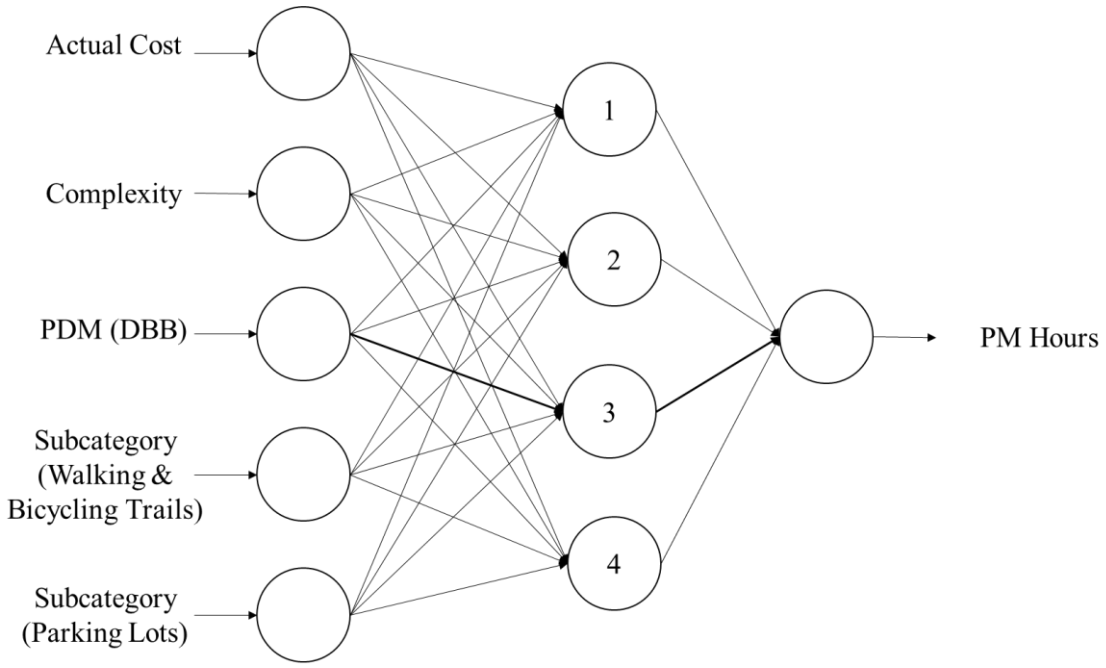


Figure 26 ANN Model of Projects <$1.4M

4.4.3 Random Forest Models

Random forest models were developed for each cluster to forecast the PM staff hours required for a project. The models were evaluated, and their performance was compared to the performance of the ANN models and the multiple linear regression models. The input variables used for each model were determined by the random forest feature selection process as opposed to the ANN and the linear regression models in which the inputs were decided using the wrapper method.

Compared to the ANN model development, the effort required to develop the random forest models was minimal. The hyperparameters that required configuration were the number of trees that need to be created and the depth of each tree. The number of trees was chosen randomly by trying 30, 50, and 100 trees. When the number of trees increases, the performance of the model increases. However, there is a point at which the improvement is minimal, but the computational time is much higher. The depth of the tree relies on the size of the data set, and the more splits it

has the more information it captures. Also, the huge depth number could lead to overfitting, and the model will not perform properly on new data. The depth range was between 1 and 10 based on the number of projects in each cluster that the model was trained on. The depth was decided by trying various depths and evaluating the model's performance. Figure 27 shows the topology of the random forest. Table 7 shows the structure of the random forest within each cluster.

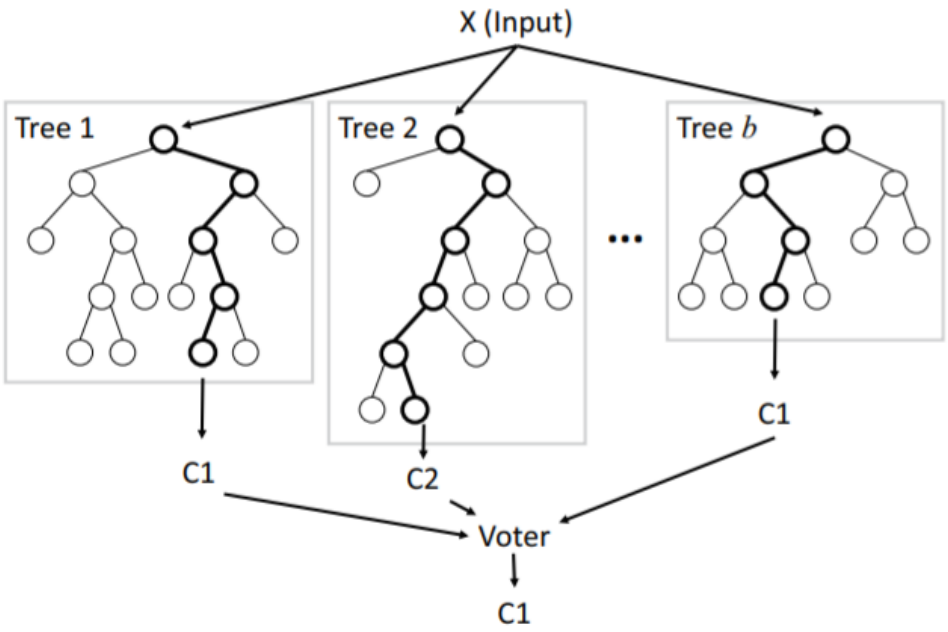


Figure 27 Random Forest Example (Nakahara et al., 2017)

Table 7 Random Forest Models Hyperparameters

Cluster	Input Variables	No. of Trees	Depth
< $1.4M	4	30	6
$1.4M–$4M	4	20	1
$4M–$12.5M	4	30	3
$12.5M–$50M	4	30	2

4.4.4 KNN Model

KNN models were developed for each cluster to forecast owner's PM staff hours required for a project. The models were evaluated, and their performance was compared to the performance of ANN, random forest, and the multiple linear regression models. The inputs required for each model were determined using the wrapper method. The KNN is a simple algorithm that predict a new example through similarity measures. In KNN, *k* refers to the number of neighbors that will be included in the majority of the voting process. Assuming that the *k* is equal to three then the prediction of a new project will be calculated from most votes of the three neighbors. The only parameter that needs to be tuned is the value of *k* which does not require a lot of effort to set compared to ANN. However, finding the best value of *k* requires some time as there is not a structured way available to follow. Choosing a smaller value of *k* can be biased towards the outliers and noisy which will impact the model results in a negative way. On the other hand, bigger value of *k* can provide better estimates with lower variance, but it is computationally expensive and could be biased sometimes. Therefore, picking a value of *k* is done by trial and error to determine which *k* will provide better performance. The value of *k* used in developing the forecasting models varied between 3 and 7 as shown in Table 8.

Table 8 KNN Parameters

Cluster	Value of *k*
< $1.4M	7
$1.4M–$4M	5
$4M–$12.5M	3
$12.5M–$50M	5

4.5 Models Evaluation

After developing the forecasting model, model validation is required to ensure that the model works properly on unseen data. Model performance can indicate whether the developed model is overfit, underfit, or generalized. There are two approaches that can be used to evaluate the performance of the developed models. The first method involves splitting the dataset into a training set and a testing set. The ideal split percentage is 70:30 or 80:20 for training and testing, respectively. The training dataset is used to train the machine learning algorithm, and the test set is used to validate the trained model. The other technique is *k*-folds cross validation, which can be used to evaluate the machine learning algorithm when the data available is limited. This method splits the input data into *k* folds between 5 to 10, depending on the size of the dataset. The model is trained using the *k*–1 folds and tested with the k^{th} fold. This process is repeated until all instances in the dataset have appeared in the training set and the test set.

Often the train/test split approach can provide biased or optimistic results when validating the model, especially with limited data. This occurs because the model misses some information from the data that was not used for training. Consequently, *k*-folds cross validation was used for model validation since it can deal with limited data because every observation in the dataset is used for training the model. Since the dataset used for training the models is considered small in size, the *k* values picked in the different models ranged between 5 and 8, depending on the number of projects in each cluster to avoid overfitting.

There were two measures used in the models' evaluation process which are the MAPE and MAE. The MAPE was the main criterion used to measure the performance vector of each forecasting model. Narula and Wellington (1977) emphasized that selecting the proper criterion is consistent with the loss function and that the MAPE is a good for evaluating the prediction model. Moreover,

Park and Stefanski (1998) indicated that in most cases the MAPE provides a meaningful performance measure when predicted values have a big range. The MAE was used as well to measure the average magnitude of the errors in each cluster. Using both metrics was beneficial in identifying the distance between the actual value and the predicted value, as well as the range of PM staff hours spent in each cluster.

Figure 28 and Figure 29 show the MAE and MAPE for the different machine learning algorithms used in each cluster to forecast the PM hours. The four different machine learning algorithms used in forecasting the PM hours had very similar results with comparable errors in each cluster. The forecasting models developed for <$1.4M cluster performed much better than the models built for the other clusters, with lower error because the number of projects used to train the models were greater than the number of projects in the other clusters. The number of projects in <$1.4M cluster is 142; the number of projects in the other clusters' ranges between 25 and 45 projects.

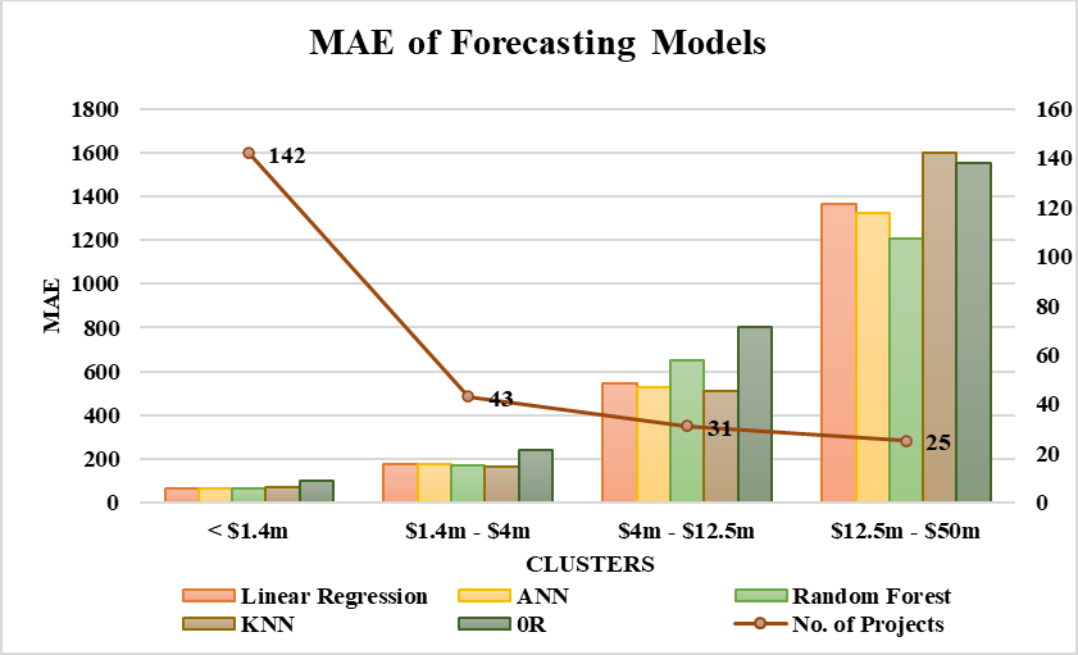


Figure 28 MAE of Forecasting Models

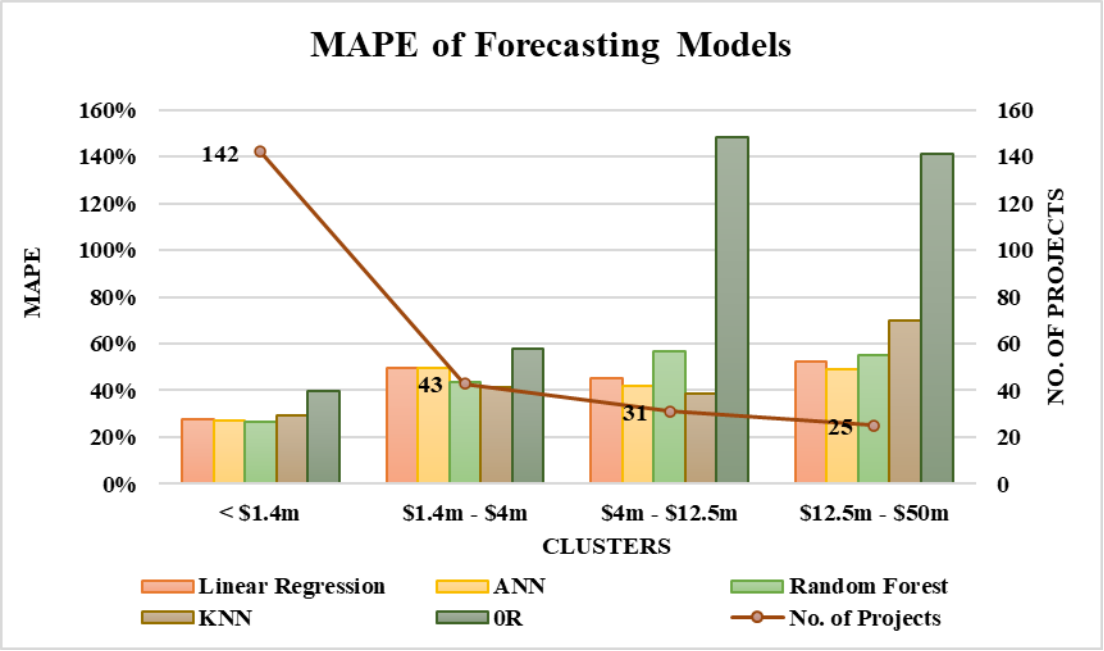


Figure 29 MAPE of Forecasting Models

The error of the forecasting models was significantly reduced after removing outliers, applying cluster analysis, and selecting features prior to the development of the forecasting model. Building one forecasting model for all cost ranges was not satisfactory and incompatible with the defined problem considering the ranges of costs and staff hours. After removing the outliers and applying feature selection techniques, the MAPE decreased from 257% to about 80%. There was a noticeable improvement; however, the error was still high and unreliable to depend on an estimate from the model. Thus, a forecasting model was developed for each cluster resulting in a noticeable improvement in the error.

The highest MAPE was about 50%, which was in the model trained using only 25 projects, which is still lower than the 80% resulted from the model developed for all projects. For the model developed for projects between $0 and $4 million, the MAPE of the model was about 40%. Also, the features from the feature selection process were not reasonable to the industry practitioners, as

the model only considered the cost of the project compared to what was implemented in this research.

ANN, linear regression, random forest and KNN model had similar MAPE and MAE. As shown in Figure 28 and Figure 29 the models with the smaller number of projects used to train them had the highest error. The MAPE of the models built for cluster $1.4M and $4M was about 50%, cluster $4M and $12.5M was about 40%, and $12.5M and $50M was about 50%. The errors of these models were high, however they performed much better than the simple average model (zero-rule). The zero-rule model which uses the dataset to calculate the average and predict the number of hours required for a given project had higher error in all clusters compared to the other models. This concludes that the proposed models can perform well in predicting the PM hours required for a project as it is better than using the average and estimates done by the industry practitioners which are ~50% accurate. Also, the models' errors are expected to be reduced when the data is collected in the proper manner as proposed in this research. Furthermore, the validation process involved assessing model results by project managers and determining if the estimates provided were reasonable. Also, their feedback was considered to improve the model's performance and to identify any drawbacks found in the model.

4.6 Bootstrapping for Range Estimation

The typical output of most machine learning algorithms is a one-point estimate. The output of the developed models is PM hours expected to be spent on an upcoming project. One-point predictions are less reliable, as the training data were scattered, the PM hours had a wide range of values, and PM hours requirements can be impacted by probabilistic events. Furthermore, one-point estimates do not provide any information regarding the level of uncertainty included in the forecasted PM

hours. In this research, bootstrapping was utilized to develop range estimates for the PM staff hour requirements.

The idea behind bootstrapping is simulating the original sampling method to generate the sampling distribution of an estimator, then bias, standard error, and confidence intervals can be calculated. The main purpose of using this method is its capability of estimating with confidence intervals, which is not an available feature in other machine learning algorithms. This method involves iterative random resampling of data with replacement. Hence, an example from the original dataset can appear zero, one, two, or more times in a bootstrap sample.

Assuming the data points are $X_1,\ldots,X_n$, the following algorithm was implemented for range estimating using empirical bootstrapping:

1. Sample with replacement from these *n* points to form a set of new observations denoted as $X_1^{*(1)},\ldots, X_n^{*(1)}$.

2. The sample procedure was repeated to create a new set of observations from the original dataset $X_1^{*(2)}, \ldots, X_n^{*(2)}$.

3. The same process was repeated for *B* times, so that the output will be *B* sets of observations. Each set of observation is called a bootstrap sample.

   a. $X_1^{*(1)}, \ldots, X_n^{*(1)}$

   b. $X_1^{*(2)}, \ldots, X_n^{*(2)}$

   c. . . .
   d. . . .
   e. . . .
   f. $X_1^{*(B)}, \ldots, X_n^{*(B)}$

4. For each bootstrap sample, ANN was fit to the *B* sets leading to *B* sets of predictions. Consequently, *B* ANNs were trained using the bootstrap samples created.

5. Finally, the confidence interval was calculated for the *B* predictions.

The *B* used in this research was equal to 30, and it was chosen based on the size of the dataset. The resampling with replacement was done on 30% of the original dataset. Using the 30 ANN trained with the bootstrap samples, 30 predictions were made for a new project. Using the mean and the standard deviation of these predictions, 60%, 75%, and 90% confidence intervals can be obtained. The confidence intervals resulted from bootstrapping includes upper and lower bounds of the PM hours expected to be spent on the project. Project managers can then evaluate the results of the prediction and select the confidence interval that they are comfortable with based on the project environment, as well as their knowledge and experience. Bootstrapping overcame the drawback of using neural network model that provides single point estimates; instead it provided project managers with range estimates for the PM hours requirements.

4.7 Conclusion

The process of data preparation was challenging as the data included noise due to the improper data collection by the owners. Different methods were used to eliminate the outliers found in the data. The cluster analysis aided in discovering the anomalies within project cost ranges.

The assumption of developing one model that can forecast any construction project was not applicable as the model had a very high error which was much worse than estimating the PM hours using an average. Therefore, the projects were grouped by actual cost and actual hours spent using k-means algorithm. Then, a forecasting model was developed for every cluster with unique inputs

obtained from the feature selection process. The nature of projects and their complexities were part of changing the features required to make a prediction from one cluster to another.

Three different machine learning algorithms – random forest, linear regression and ANN – were evaluated to determine which performed better. The MAPE and MAE of the model were within the same range; however, the ANN was picked for future use due to its ability in handling noisy and nonlinear data and dealing with complex problems.

The model developed for the projects that are less than $1.4 million has lower error compared to the other models since the number of projects used to train that model was larger than that of other projects. Even though, the model has ±27% error, which is better than the error of the models utilized by the industry practitioners (~40%), the error is still high. This is caused by the lack of projects used for training the models, the inconsistency of collecting and tracking projects information such as the spent LHRs, and the lack of available project information. The accuracy of the models is expected to increase when the project data is collected based on the proposed data acquisition model, and by increasing the number of projects to train the neural network model.

Bootstrapping provides a range of PM hours predicted by the neural network. The method is easy to implement, and it can provide project managers with confidence intervals regarding their estimates.

Chapter 5: Decision Support System for PM Resource Forecasting

5.1 Introduction

A decision support system computer application for estimating owner's PM staff requirements for construction projects was developed. The application consists of two modules; the first module is for tracking and storing project data, and the second module is the analytical model to do the forecasting process. Both modules are integrated so that the user does not have to worry about handling two applications. One thing that needed to be considered while developing the application was avoiding duplication work. Since some of the owners were using ERP systems and PM systems, the data from these systems can be imported into the application through predefined Excel sheets. Also, the user interface should be user friendly and provide project managers with the required information, reports, or graphs in a simple way. Figure 30 illustrates the application framework including the two modules and the input required to forecast the PM hours.

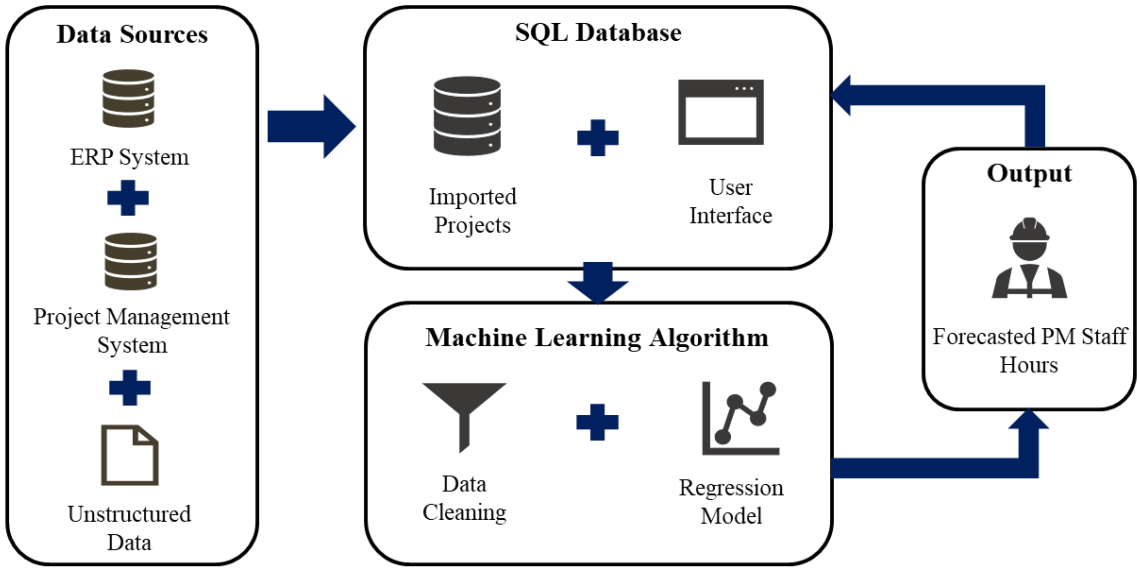


Figure 30 Decision Support System Framework

5.2 Objectives

The objectives of developing the computer application included:

1. Improving the data collection process.

2. Providing a user-friendly, flexible database for entering and storing projects, portfolios, and resources information.

3. Providing a framework that to help future analyses and evaluating owner's performance.

4. Improving the PM staff estimation process and providing reliable estimates from historical data.

5.3 Current Estimating Systems

The current systems used for estimating the PM staff requirements are speadsheets. Knowledge and experience from project managers was the main the approach used for estimating the resources. Also, one of the departments used historical data for some estimates; however, this approach provided a large error. The details of the approaches and the systems adopted by the owners can be found in Chapter 3.

5.4 System Design and Components

The decision support system is composed of two modules: a database module for storing and tracking detailed data collection, and an analytical module that includes the ANN model to forecast PM staff requirements from the data stored in the database or from an uploaded CSV file.

5.4.1 Database Module

The database module was developed using Microsoft SQL Server to aid owners in collecting, tracking, and storing detailed information. It is possible to enter new data manually or import data from other systems through formatted Excel sheets. The ERD diagram shown in Figure 4 in Chapter 3 includes all the entities involved in the development of this module. This module collects project information towards the project phases. Moreover, some predefined values used in the database can be modified and entered by the user based on their preference, e.g. the resource types working for the departments, PDM, project categories and subcategories, and project phases. However, any interpretation of the predefined values will cause retraining and revalidation of the neural network model. The functions of this module are to add and edit new data, produce reports on collected information, and provide a user interface to for the analytical module.

The home page that allows the users to access different fields in which they can edit and add information related to these fields as shown in Figure 31. The “General Setup” involves all the predefined attributes of each entity that can be modified by the user. “Owner Companies” includes the owner, business units, and departments that manage the projects stored in the “Projects” section. “Portfolios/Programs” includes the portfolios which consist of multiple programs, and the programs which comprise several projects. “Projects” includes adding and editing projects and their attributes such as actual cost, budget, planned duration, PDM, etc. It also includes the project phases such as initiation, planning, design, construction and warranty. The phases include some attributes such as budget, cost, and duration as well. Figure 32 provides an example of the predefined attributes such as the PDM and how the user can add, edit or remove any of the values.

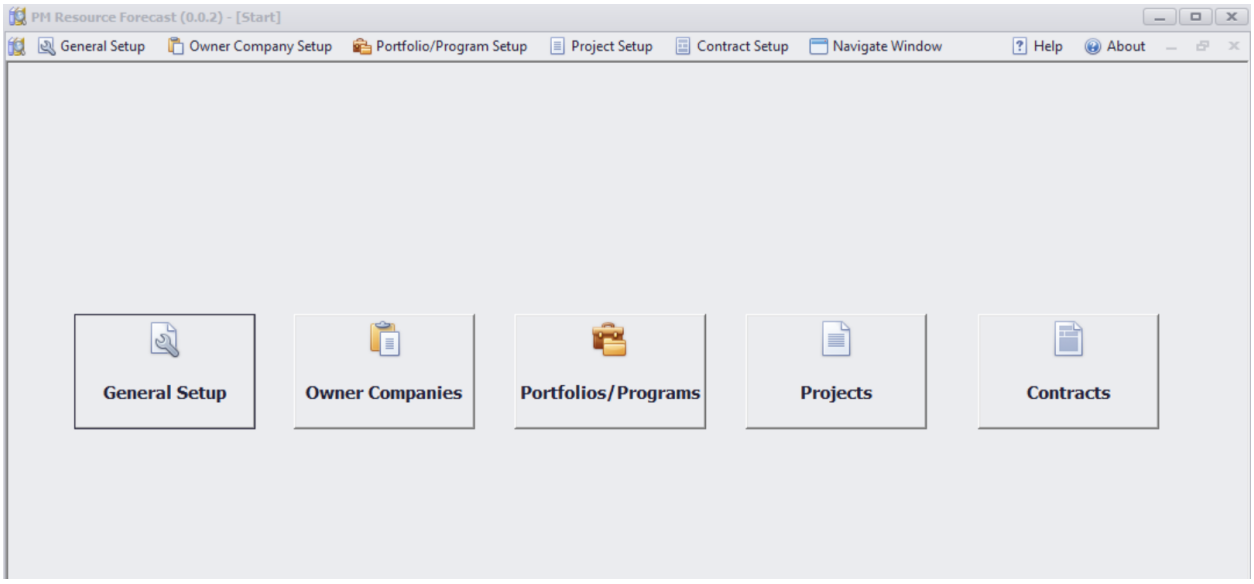


Figure 31 Home Page of the Database Module

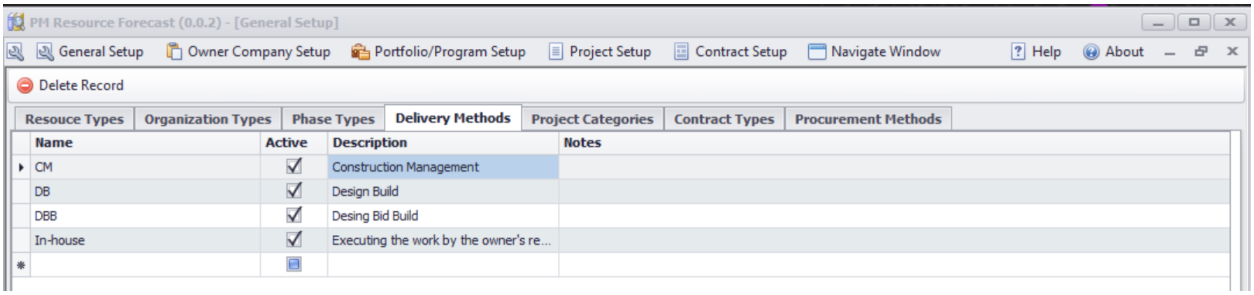


Figure 32 Example for Modification of Predefined Values in General Setup

5.4.1.1 Database Tables

The ERD shown in Figure 4 involves multiple entities and their relationship with each other. Each entity includes attributes that describe the entity, for example the project entity has a project name and a project ID as attributes. Since the attributes of the entities are missing in the ERD for simplicity purposes, the database tables are shown in Figure 33 which include the entities' attributes.

owners

owner_id
owner_name (AK1) org_type_id (FK) active notes (O)

Owners Table

departments

dept_id owner_id (FK,FK,AK1)
dept_name (AK1) parent_id (O) (FK) resources active notes (O)

Departments Table

portfolios

portfolio_id
owner_id (FK) dept_id (FK,AK1) portfolio_name (AK1) active notes (O)

Portfolios Table

programs

program_id portfolio_id (FK,AK1)
prog_name (AK1) active notes (O)

Programs Table

resource_types

resource_id
resource_name (AK1) resource_description (O) active notes (O)

Resources Table

contracts

contract_id
cont__name (AK1) proj_id (FK) contract_type_id (O) (FK) proc_method_id (O) (FK) budget (O) commited (O) active notes (O)

Contracts Table

proj_phases_dept_resources

proj_id (FK) phase_id (FK) resource_id (FK) owner_id (FK) dept_id (FK)
pred_qty (O) notes (O)

Project Phase and Resources Relationship Table

proj_phases

phase_id proj_id (FK,AK1)
phase_name (O) phase_type_id (FK,AK1) phase_budget (O) phase_cost (O) phase_plan_duration (O) phase_duration (O) owner_hours (O) notes (O) active completed

Project and Project Phase Relationship Table

proj_resources

proj_id (FK) resource_id (FK)
resource_hrs (O) req_hrs (O) ratio (O) notes (O)

Project and Resources Relationship Table

projects

proj_id
owner_id (FK,AK1) proj_name (AK1) program_id (O) (FK) method_id (O) (FK) category_id (O) (FK,FK) sub_cat_id (O) (FK) proj_type (O) proj_complexity (O) location (O) public_interest budget (O) cost (O) start_date (O) finish_date (O) duration (O) phase_type_id (FK) project_stage (O) notes (O) active approved_budget (O) act_start_date (O) act_finish_date (O) rev_budget (O) base_duration (O) act_duration (O) variance (O) completed

Project Table

Figure 33 Database Tables

5.4.2 Analytical Module

The analytical module includes two parts, and it is responsible for performing the PM resource forecasting. The first part includes the training portion of the neural network that was developed based on the historical projects collected in this research. The second part involves the forecasting portion that uses the trained neural network to estimate the PM hours required for a new project.

All project information is collected in the database module; however, the forecasting module requires certain inputs to do the forecasting. As previously mentioned, a forecasting model was developed for each project's cost range and each requires different inputs. One of the trained neural networks will be used to forecast the new project PM staff hours based on the budget, and the selection of the neural network is made when the user picks the project budget. For simplicity, only developers will have access to retrain the models or modify the neural networks structure and hyperparameters. To predict the PM hours, the user fills in the required inputs. The following rules were set to allow a proper training and forecasting of the neural networks:

- Resource forecasting is performed for a single project with status equals "Not Completed".
- Forecasting is done for PM resource types only.
- Projects with status equals to "Completed" are the only ones used to train the models.
- Projects with PM staff hours less than 100 LHRs are excluded from training part.
- Projects with Actual Cost equals to 0 or no value are not included in the training part.

The model provides the user with multiple outputs from the analytical module, including the PM hours predicted by the neural network model, the confidence intervals of the prediction, and a distribution graph for the estimate. The analytical module functions are training neural networks using the data stored in the database module; forecasting the PM hours requirements for a new project; and providing a range estimate for the predicted PM hours.

This module will provide project managers with insights regarding the PM hours required for a given project. For example, the outputs of a new project, that has a budget of $1M and its features

are defined in the database including PDM, subcategory and complexity, are single-point estimate for PM hours, 60%, 75% and 90% confidence levels and PDF for visualization as shown in Figure 34 and Table 9.

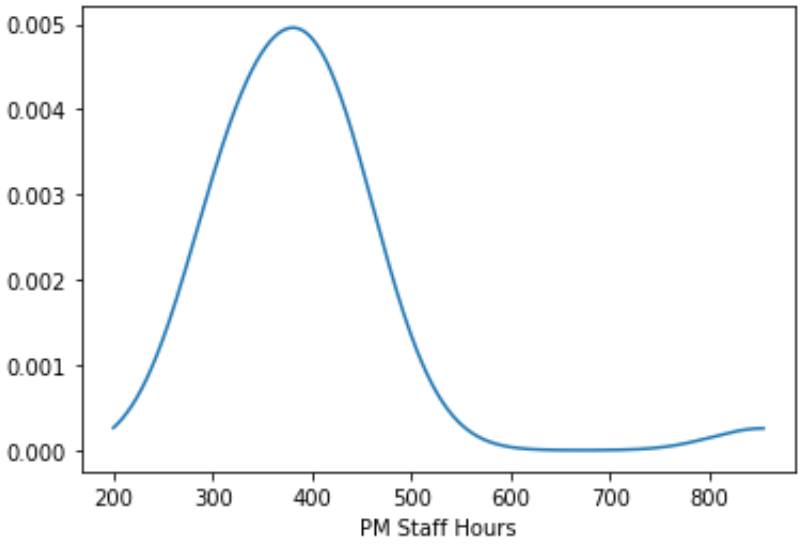


Figure 34 PM Staff Hours Distribution for a Project

Table 9 Analytical Module Outputs

Output	Result
Single-point Prediction	392
90% Confidence Level	[368, 415]
75% Confidence Level	[380, 404]
60% Confidence Level	[387, 397]

5.5 Proposed Reports

Reports are used to evaluate business problems and communicate information that includes results and analyses of data. Moreover, they can be used to support crucial decisions taken by top management. The data collected through the proposed data acquisition model will aid project managers in taking corrective action in future projects and making important decisions during the project.

A report that can be exported from the application will help in evaluating the workload in the upcoming year against the available PM staff on a departmental level. The graph in Figure 35 shows the forecasted hours required to manage projects in the upcoming year, taking into account projects that are currently in progress and those expected to start within the year. The graphs show the months with a lack of PM resources to handle the workload which might lead to reallocating work or resources. Also, this graph can aid the decisions of project managers in hiring new staff.

Another report that can be pulled out of from the collected and tracked projects information is the forecasted LHRs and the actual LHRs spent on each project. The graph in Figure 36 shows actual and the forecasted LHRs in different projects within a department. This report will help to ensure that the forecasting models are working properly and that the errors of the estimates are within an acceptable range. If the estimates are inaccurate it means that the forecasting models need to be retrained and validated using the new store records in the database.

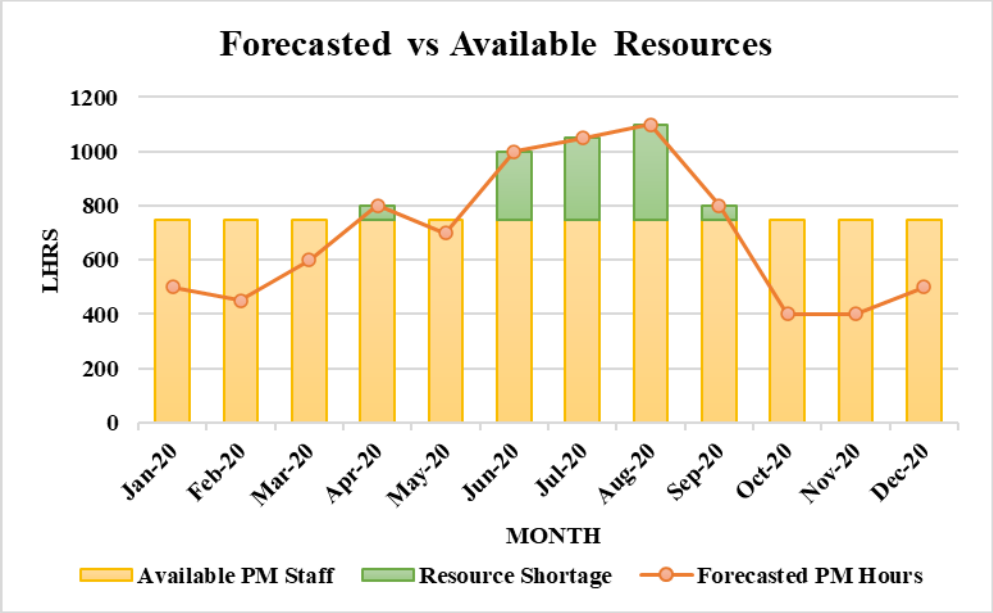


Figure 35 Departmental PM Forecast vs Available

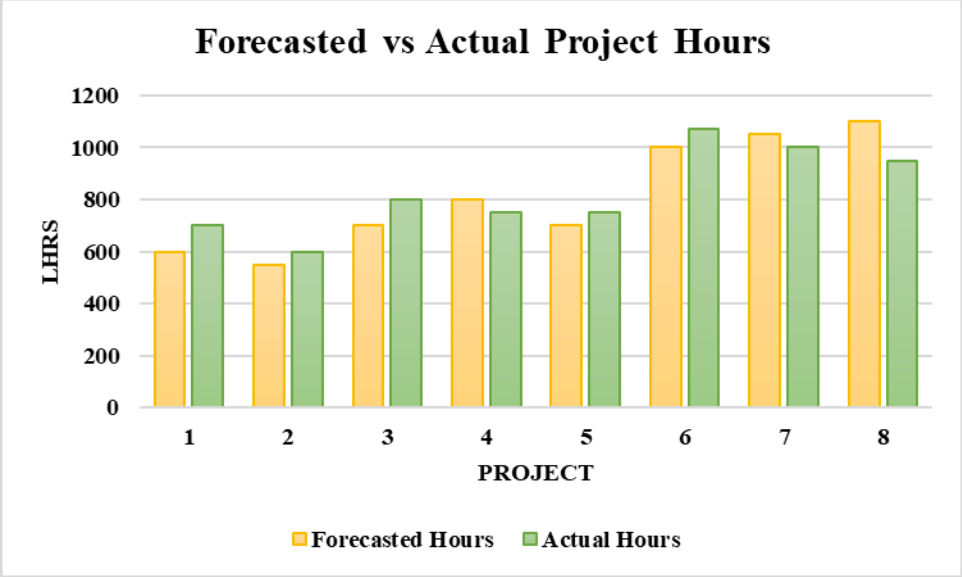


Figure 36 Forecasted vs Actuals Projects Hours

Chapter 6: Final Discussion

The objective of the research was to develop a model that can forecast owner's PM resources in hours for a given construction project. The objective was achieved by creating multiple ANN models for different project cost ranges. Yet, the error of the models was high and varied from each other due to the inconsistency found in the dataset and the limited number of projects used to train the models.

First, project factors affecting the staff requirements were investigated considering the previous studies carried out by researchers and the industry practices. A preliminary project attributes list was conducted which included literature and industry factors. Historical project data were collected based on the defined project attributes. This list consisted of objective and subjective data, and it included the following attributes: actual cost and duration; spent PM hours; project complexity, category (buildings, industrial, or infrastructure), subcategory (e.g. business, recreation, parkade, etc.), and type (renovation or new); PDM; public engagement, project field type (green or brown) and location.

Second, the data collection process was challenging because much of the outlined project attributes were not available or tracked by the owners. A portion of projects information was collected from the PM systems, ERP systems, and status reports, while the other portion was collected manually through discussions with the project managers who oversaw the projects. This caused some limitations in the collected dataset. Project managers were relying on their memory when recalling the PDM and level of public engagement. Although, they were provided with a framework on how to select the level of complexity of a project, their selection could be biased as based their ratings on their knowledge and experience. Sometimes, project attributes such as the actual duration and the spent PM hours were not tracked by owners. Thus, the duration was calculated from the first

and last transactions of the project, and the hours were computed from multiplying the number of PM staff who worked on the project by the number of hours per month and number of months.

Next, the data were prepared and of data mining techniques were applied to obtain useful knowledge and develop a forecasting model. The assumption of developing one model for any construction project was invalid due to a very high error that was worse than estimating the PM hours using averages. Therefore, clustering analysis was applied on the dataset to group projects by actual cost and actual hours spent. Then, forecasting model was developed for each cluster after defining the required inputs for each model using feature selection techniques. Table 10 shows the inputs required for each developed model. Three different machine learning algorithms – random forest, linear regression, KNN and ANN – were evaluated to determine which performed better. The MAPE and MAE of the models were within the same range; however, ANN was picked for future use due to its ability in handling noisy and nonlinear data and dealing with complex problems.

Table 10 Model Inputs

Model	Model Inputs
< $1.4M	Actual cost Complexity PDM (DBB) Subcategory (Walking & Bicycling Trails, Parking Lots)
$1.4M–$4M	Actual cost Complexity Duration Subcategory (Roads)
$4M–$12.5M	Actual cost' Subcategory (High Security Facilities, Demolition)
$12.5M–$50M	Public interest Duration PDM (CM) Subcategory (Business, Roads)

The ANN models were evaluated, and the results are shown in Table 11. The model developed for projects that less than $1.4 millions has lower error compared to the other models due to the size of the training set. That model has an error of ± 27%, which is better than the error of models currently in use by industry practitioners (~40%), but the error is still high. This was caused by the limited number of records available in the dataset, the inconsistency of collecting and tracking projects information such as the spent LHRs, and the lack of available projects information. The

accuracy is expected to increase when the project data is collected based on the proposed data acquisition model and the size of the dataset increases.

Table 11 ANN Models Results

Cluster	ANN	
	MAPE	***MAE***
< $1.4M	26.9%	+/-65 LHRs
$1.4M–$4M	49.7%	+/-177 LHRs
$4M–$12.5M	41.7%	+/-526 LHRs
$12.5M–$50M	49.1%	+/-1323 LHRs

The output of the developed models is the PM hours expected to be spent on an upcoming project as is a single point estimate, which is considered unreliable in the industry practices. Consequently, bootstrapping was used to make a prediction with a confidence interval to support project managers in making their decisions regarding the staff requirements.

Finally, a computer application consisting of two modules was developed. The first module is the data acquisition system, developed in SQL database and capable of tracking and storing owner's data. This module was structured for easy data collection and future analyses, not limited to estimating PM resources. The second module is the analytical model that includes the neural network model and estimates the PM staff requirements from stored data.

This research has proposed beneficial contributions to both academia and industry. The academic contributions involve combining the conducted research and industry practices to identify significant project features affecting staff requirements and apply data mining techniques on a

practical problem to forecast owners PM staff requirements for any construction project. The industrial contributions include developing a computer application that has the capability of tracking and storing projects data in a proper format and forecasting owner's PM hours required for a project from the stored data. Also, the application will aid in the identifying the current available resources and support hiring decisions.

The developed decision support system has few limitations. The records used to train the model are considered small, and data mining requires a large number of records for better accuracy. The data used to train the models was inconsistent and might include additional anomalies that were not identified. For instance, few projects might have over allocated hours since they were under budget during execution and vice versa. Another example is tracking the hours for some, but not all, of a project due to the unavailability of a proper tracking system. Thus, the models need to be revalidated after a larger number of projects is acquired to ensure the best performance and a higher prediction accuracy.

Recommendations for future research includes expanding the model to forecast owner's PM requirements on a phase level because the workload of PM staff fluctuates during the project execution. This will allow reallocation of resources during project execution more resources are required to manage the job. This part was already included in the data acquisition model so it can be collected and stored for future use. The model could also be generalized to forecast PM resources for contractors, which might require collecting additional information. The model can be expanded to forecast different types of resources, not just the PM resources; however, contractors and consultants will be more concerned with that part than owners.

References

Ahadzie, D. K., Proverbs, D. G., & Olomolaiye, P. O. (2008). Model for predicting the performance of project managers at the construction phase of mass house building projects. *Journal of Construction Engineering and Management*, *134*(8), 618–629. https://doi.org/10.1061/(ASCE)0733-9364(2008)134:8(618)

Ahiaga Dagbui, D. D., & Smith, S. D. (2014). Rethinking construction cost overruns: Cognition, learning and estimation. *Journal of Financial Management of Property and Construction*, *19*(1), 38–54. https://doi.org/10.1108/JFMPC-06-2013-0027

Amiri, M., Ardeshir, A., Fazel Zarandi, M. H., & Soltanaghaei, E. (2016). Pattern extraction for high-risk accidents in the construction industry: a data-mining approach. *International Journal of Injury Control and Safety Promotion*, *23*(3), 264–276. https://doi.org/10.1080/17457300.2015.1032979

André, M., Baldoquín, M. G., & Acuña, S. T. (2011). Formal model for assigning human resources to teams in software projects. *Information and Software Technology*, *53*(3), 259–275. https://doi.org/10.1016/j.infsof.2010.11.011

Art Chaovalitwongse, W., Wang, W., Williams, T. P., & Chaovalitwongse, P. (2012). Data Mining Framework to Optimize the Bid Selection Policy for Competitively Bid Highway Construction Projects. *Journal of Construction Engineering and Management*, *138*(2), 277–286. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000386

Ballesteros-Pérez, P., González-Cruz, M. C., & Fernández-Diego, M. (2012). Human resource allocation management in multiple projects using sociometric techniques. *International*

Journal of Project Management, *30*(8), 901–913. https://doi.org/10.1016/j.ijproman.2012.02.005

Bell, L. C., & Brandenburg, S. G. (2003). Forecasting construction staffing for transportation agencies. *Journal of Management in Engineering*, *19*(3), 116–120. https://doi.org/10.1061/(ASCE)0742-597X(2003)19:3(116)

Bhatia, P. (2019). *Data Mining and Data Warehousing Principles and Practical Techniques*. Cambridge University Press. https://doi.org/10.1017/9781108635592

Blichfeldt, B. S., & Eskerod, P. (2008). Project portfolio management - There's more to it than what management enacts. *International Journal of Project Management*, *26*(4), 357–365. https://doi.org/10.1016/j.ijproman.2007.06.004

Bonaccorso, G. (2018). *Machine Learning Algorithms - Second Edition*. Packt Publishing. https://learning.oreilly.com/library/view/python-machine-learning/9781787125933/ch03s06.html

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/bf00058655

Carney, J. G., Cunningham, P., & Bhagwan, U. (1999). Confidence and prediction intervals for neural network ensembles. *Proceedings of the International Joint Conference on Neural Networks*, *2*, 1215–1218. https://doi.org/10.1109/ijcnn.1999.831133

Chen, J. H., Yang, L. R., Chen, W. H., & Chang, C. K. (2008). Case-based allocation of onsite supervisory manpower for construction projects. *Construction Management and Economics*, *26*(8), 805–814. https://doi.org/10.1080/01446190802014778

Dabirian, S., Abbaspour, S., Khanzadi, M., & Ahmadi, M. (2019). Dynamic modelling of human resource allocation in construction projects. *International Journal of Construction Management*, *3599*(May). https://doi.org/10.1080/15623599.2019.1616411

Dey, P. K. (2012). Project risk management using multiple criteria decision-making technique and decision tree analysis: A case study of Indian oil refinery. *Production Planning and Control*, *23*(12), 903–921. https://doi.org/10.1080/09537287.2011.586379

Engwall, M., & Jerbrant, A. (2003). The resource allocation syndrome: The prime challenge of multi-project management? *International Journal of Project Management*, *21*(6), 403–409. https://doi.org/10.1016/S0263-7863(02)00113-8

Fini, A. A. F., Akbarnezhad, A., Rashidi, T. H., & Waller, S. T. (2018). Dynamic programming approach toward optimization of workforce planning decisions. *Journal of Construction Engineering and Management*, *144*(2), 1–14. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001434

Franke, J., & Neumann, M. H. (2000). Bootstrapping neural networks. *Neural Computation*, *12*(8), 1929–1949. https://doi.org/10.1162/089976600300015204

Gorunescu, F. (2011). *Data Mining: Concepts, Models and Techniques*. Springer-Verlag. https://doi.org/10.1007/978-3-642-19721-5

Hammad, A., Abourizk, S., & Mohamed, Y. (2014). Application of KDD techniques to extract useful knowledge from labor resources data in industrial construction projects. *Journal of Management in Engineering*, *30*(6), 1–10. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000280

Han, J., Kamber, M., & Pei, J. (2012). Data Mining. In *Data Mining* (Third Edit). Elsevier. https://doi.org/10.1016/C2009-0-61819-5

He, Q., Luo, L., Hu, Y., & Chan, A. P. C. (2015). Measuring the complexity of mega construction projects in China-A fuzzy analytic network process analysis. *International Journal of Project Management*, *33*(3), 549–563. https://doi.org/10.1016/j.ijproman.2014.07.009

Hendriks, M., Voeten, B., & Kroep, L. (1999). Human resource allocation in a multi-project R&D environment. *International Journal of Project Management*, *17*(3), 181–188. https://doi.org/10.1016/s0263-7863(98)00026-x

Ho, P. H. K. (2010). Forecasting construction manpower demand by gray model. *Journal of Construction Engineering and Management*, *136*(12), 1299–1305. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000238

Holmes, H., Garcia-Taengua, E., & Fuentes, R. (2019). Meta-analysis of ground movements associated with deep excavations using a data mining approach. *Journal of Rock Mechanics and Geotechnical Engineering*, *11*(2), 409–416. https://doi.org/10.1016/j.jrmge.2018.12.006

Huang, C. H., & Hsieh, S. H. (2020). Predicting BIM labor cost with random forest and simple linear regression. *Automation in Construction*, *118*. https://doi.org/10.1016/j.autcon.2020.103280

Jun; Khaled El-Rayes, D. H. (2011). Multiobjective Optimization of Resource Leveling and Allocation during Construction Scheduling. *Journal of Construction Engineering and Management*, *137*(12), 1080–1088. https://doi.org/10.1061/(ASCE)CO .1943-7862.0000368

Kang, K., & Ryu, H. (2019). Predicting types of occupational accidents at construction sites in Korea using random forest model. *Safety Science*, *120*, 226–236. https://doi.org/10.1016/j.ssci.2019.06.034

Lajbcygier, P. R., & Connor, J. T. (1997). Improved option pricing using artificial neural networks and bootstrap methods. *International Journal of Neural Systems*, *8*(4), 457–471. https://doi.org/10.1142/S0129065797000446

Lee, M.-J., Hanna, A. S., & Loh, W.-Y. (2004). Decision Tree Approach to Classify and Quantify Cumulative Impact of Change Orders on Productivity. *Journal of Computing in Civil Engineering*, *18*(2), 132–144. https://doi.org/10.1061/~ASCE!0887-3801~2004!18:2~132!

Leung, M., Yee-Shan, C., & Paul, O. (2008). Impact of Stress on the Performance of Construction Project Managers. *Journal of Construction Engineering and Management*, *134*(8), 644–652. https://doi.org/10.1061/(ASCE)0733-9364(2008)134

Lin, K. L. (2011). Human resource allocation for remote construction projects. *Journal of Management in Engineering*, *27*(1), 13–20. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000032

Lins, I. D., Droguett, E. L., Moura, M. D. C., Zio, E., & Jacinto, C. M. (2015). Computing confidence and prediction intervals of industrial equipment degradation by bootstrapped

support vector regression. *Reliability Engineering and System Safety*, *137*, 120–128. https://doi.org/10.1016/j.ress.2015.01.007

Lowe, D. J., Emsley, M. W., & Harding, A. (2006). Predicting Construction Cost Using Multiple Regression Techniques. *Journal of Construction Engineering and Management*, *132*(7), 750–758. https://doi.org/10.1061/(ASCE)0733-9364(2006)132:7(750)

Lu, M., AbouRizk, S. M., & Hermann, U. H. (2000). Estimating Labor Productivity Using Probability Inference Neural Network. *Journal of Computing in Civil Engineering*, *14*(4), 241–248. https://doi.org/10.1061/(ASCE)0887-3801(2000)14:4(241)

Lu, Y., Luo, L., Wang, H., Le, Y., & Shi, Q. (2015). Measurement model of project complexity for large-scale projects from task and organization perspective. *International Journal of Project Management*, *33*(3), 610–622. https://doi.org/10.1016/j.ijproman.2014.12.005

Martinsuo, M., & Lehtonen, P. (2007). Role of single-project management in achieving portfolio management efficiency. *International Journal of Project Management*, *25*(1), 56–65. https://doi.org/10.1016/j.ijproman.2006.04.002

Meehan, R. H., & Ahmed, S. B. (1990). Forecasting Human Resources Requirements: A Demand Model. *Human Resource Planning*, *13*(4), 297–307. http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=7700712&site=ehost-live

Meskendahl, S. (2010). *The influence of business strategy on project portfolio management and its success-A conceptual framework*. https://doi.org/10.1016/j.ijproman.2010.06.007

Moreno, J. L. (1960). *The sociometry reader,*. Free Press.

Mourya, S. K., & Gupta, S. (2012). *Data Mining and Data Warehousing*. Alpha Science International Ltd. http://ebookcentral.proquest.com/lib/ualberta/detail.action?docID=5218420

Nakahara, H., Jinguji, A., Sato, S., & Sasao, T. (2017). A Random Forest Using a Multi-valued Decision Diagram on an FPGA. *Proceedings of The International Symposium on Multiple-Valued Logic*, 266–271. https://doi.org/10.1109/ISMVL.2017.40

Narula, S. C., & Wellington, J. F. (1977). Prediction, linear regression and the minimum sum of relative errors. *Technometrics*, *19*(2), 185–190. https://doi.org/10.1080/00401706.1977.10489526

Othman, M., Bhuiyan, N., & Gouw, G. (2011). A New Approach to Workforce Planning. *World Academy of Science, Engineering and Technology*, *76*(4), 804–811. https://waset.org/journals/waset/v52/v52-159.pdf

Paass, G. (1992). Assessing and Improving Neural Network Predictions by the Bootstrap Algorithm. *Advances in Neural Information Processing Systems*, 196–203.

Papadopoulos, G., Edwards, P. J., & Murray, A. F. (2001). Confidence estimation methods for neural networks: A practical comparison. *IEEE Transactions on Neural Networks*, *12*(6), 1278–1287. https://doi.org/10.1109/72.963764

Parisi Kern, A., Ferreira Dias, M., Piva Kulakowski, M., & Paulo Gomes, L. (2015). Waste generated in high-rise buildings construction: A quantification model based on statistical multiple regression. *Waste Management*, *39*(2015), 35–44. https://doi.org/10.1016/j.wasman.2015.01.043

Park, H., & Stefanski, L. A. (1998). Relative-error prediction. In *Statistics & Probability Letters* (Vol. 40).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. In *Journal of Machine Learning Research* (Vol. 12). http://scikit-learn.sourceforge.net.

Persad, K. R., O'Connor, J. T., & Varghese, K. (1995). Forecastng engineering manpower requirements for highway preconstruction activities. *Journal of Management in Engineering*, *11*(3), 41–47. https://doi.org/10.1061/(ASCE)0742-597X(1995)11:3(41)

Proverbs, D. G., Holt, G. D., & Olomolaiye, P. O. (1999). A method for estimating labour requirements and costs for international construction projects at inception. *Building and Environment*, *34*(1), 43–48. https://doi.org/10.1016/s0360-1323(97)00064-4

Raschka, S. (2017). *Python Machine Learning* (Second edi). Wiley. https://learning.oreilly.com/library/view/machine-learning-algorithms/9781789347999/46ad22e9-6630-4611-8463-fe1657873b1f.xhtml

Silva, L. C. e, & Costa, A. P. C. S. (2013). Decision model for allocating human resources in information system projects. *International Journal of Project Management*, *31*(1), 100–108. https://doi.org/10.1016/j.ijproman.2012.06.008

Sing, C. P., Love, P. E. D., & Tam, C. M. (2014). Forecasting the demand and supply of technicians in the construction industry. *Journal of Management in Engineering*, *30*(3), 1–9. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000227

Sing, C. P., Love, P. E. D., & Tarn, C. M. (2012). Multiplier model for forecasting manpower demand. *Journal of Construction Engineering and Management*, *138*(10), 1161–1168. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000529

Sonmez, R. (2011). Range estimation of construction costs using neural networks with bootstrap prediction intervals. *Expert Systems with Applications*, *38*(8), 9913–9917. https://doi.org/10.1016/j.eswa.2011.02.042

Vujičić, T., & Matijevi, T. (2016). Comparative Analysis of Methods for Determining Number of Hidden Neurons in Artificial Neural Network. *Central European Conference on Information and Intelligent Systems*, 219–223.

Wang, R., Asghari, V., Hsu, S. C., Lee, C. J., & Chen, J. H. (2020). Detecting corporate misconduct through random forest in China's construction industry. *Journal of Cleaner Production*, *268*. https://doi.org/10.1016/j.jclepro.2020.122266

Witten, I., Frank, E., Hall, M. A., & Pal, C. J. (2011). Data Mining: Practical Machine Learning Tools and Techniques. In *Data Mining: Practical Machine Learning Tools and Techniques* (Fourth Edi). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-374856-0.00026-2

Wong, J. M. W., Chan, A. P. c., & Chiang, Y. H. (2011). Construction manpower demand forecasting: A comparative study of univariate time series, multiple regression and econometric modelling techniques. *Engineering, Construction and Architectural Management*, *18*(1), 7–29. https://doi.org/10.1108/09699981111098667

Wong, J. M. W., Chan, A. P. C., & Chiang, Y. H. (2007). Forecasting construction manpower demand: A vector error correction model. *Building and Environment*, *42*(8), 3030–3041. https://doi.org/10.1016/j.buildenv.2006.07.024

Wong, J. M. W., Chan, A. P. C., & Chiang, Y. H. (2008). Modeling and forecasting construction labor demand: Multivariate analysis. *Journal of Construction Engineering and Management*, *134*(9), 664–672. https://doi.org/10.1061/(ASCE)0733-9364(2008)134:9(664)

Wu, M. C., & Sun, S. H. (2006). A project scheduling and staff assignment model considering learning effect. *International Journal of Advanced Manufacturing Technology*, *28*(11–12), 1190–1195. https://doi.org/10.1007/s00170-004-2465-0

Yang, W. J., & Kim, Y. S. (2019). Manpower Allocation Model for Construction Site Office Engineers based on Inherent Technical Risks. *KSCE Journal of Civil Engineering*, *23*(3), 947–957. https://doi.org/10.1007/s12205-019-0663-4

Zhang, J. (1999). Developing robust non-linear models through bootstrap aggregated neural networks. *Neurocomputing*, *25*(1–3), 93–113. https://doi.org/10.1016/S0925-2312(99)00054-5

Zio, E. (2006). A study of the bootstrap method for estimating the accuracy of artificial neural networks in predicting nuclear transient processes. *IEEE Transactions on Nuclear Science*, *53*(3), 1460–1478. https://doi.org/10.1109/TNS.2006.871662