

A Forecasting Model for Labor Resources in Construction Projects Using Machine Learning

by

Hamidreza Mohammadhosseinzadeh Golabchi

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

© Hamidreza Mohammadhosseinzadeh Golabchi, 2021

Abstract

Considering the high rates of labor resources in construction projects clearly indicates the importance of appropriate labor resource management methods. Accurate labor resource allocation is a substantial step towards successful labor resource management. With the recent developments in the area of artificial intelligence and machine learning, these technologies can potentially be adopted to develop prediction models. This research aims to combine the benefits of artificial intelligence and historical data of previous projects to identify the significant factors affecting the labor resource requirements and to develop an efficient predictive model to analyze and learn from the past construction projects in order to have a precise estimate of required labor hours for upcoming projects.

The research involves collecting and analyzing historical data, investigating current industry practices in labor resource estimation, and implementing machine learning algorithms to predict required labor hours for various resources in construction projects. Also, this study explores the key factors impacting the needed labor resources by combining the literature review and industry practices. Furthermore, this thesis offers a neural network model which can forecast the required labor hours for a construction work package by utilizing the historical data. The proposed model provides the estimation of the required labor hours for each day of the work package. The developed model aids project managers in labor resource allocation in construction projects and provides them a precise insight to perform a decent labor resource management.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisor Dr. Ahmed Hammad for his continuous support and consistent encouragement during my studies.

I am also thankful to Maria Al-Hussein for her supportive assistance and many thanks to Trever Harrington for providing invaluable and professional feedback.

I owe my deepest gratitude to my parents for their never-ending assistance and motivation during every stage of my life.

Finally, I would like to thank my brilliant brother and his kind wife for their ongoing encouragement. Without their support, this could not have been possible.

Table of Contents

Chapter 1 Introduction	1
1.1 Background.....	1
1.2 Research Objectives.....	2
1.3 Research Methodology	3
1.4 Thesis Organization	4
Chapter 2 Literature Review	5
2.1 Human Resource Management (HRM) in Construction.....	5
2.2 Factors Affecting HR Demand in Construction Projects	11
2.3 Applying Machine Learning Algorithms in Construction Projects	14
2.4 Research Gaps.....	20
Chapter 3 Methodology for Forecasting Labor Resources	22
3.1 Introduction.....	22
3.2 Significant Key Factors Affecting Required Labor Resources.....	24
3.2.1 Industry Practices in Labor Resource Estimating	24
3.2.2 Selection of Factors.....	26
3.3 Data Acquisition Model for Data Collection	28
3.3.1 Entity Relationship Diagram (ERD).....	29
3.3.2 Attributes Required for Forecasting Labor Resources.....	32
3.4 Data Collection Process	33
3.4.1 Essential Attributes for the Forecasting Model.....	34
3.4.2 Data Sources	34
3.4.3 Missing Information.....	35
3.4.4 Data Transformation and Linking.....	35
3.4.5 Data Understanding and Visualization.....	36
3.4.6 Deficiencies in the Collected Dataset	40

Chapter 4 Development of Forecasting Model for Labor Resources.....	42
4.1 Introduction.....	42
4.2 Data Preprocessing.....	43
4.2.1. Missing Values.....	44
4.2.2. Cluster Analysis and Outliers Detection	44
4.2.3. Nominal Features	47
4.2.4. Normalizing & Data Splitting	48
4.3 Feature Selection.....	50
4.4 Forecasting Models.....	52
4.4.1. Baseline Model	53
4.4.2. Linear Model.....	53
4.4.3. Recurrent Neural Network Model.....	54
4.5 Performance Evaluation.....	57
4.6 Summary.....	59
Chapter 5 Implementation of the Labor Resource Forecasting Model	61
5.1 Introduction.....	61
5.2 Application Components.....	64
Chapter 6 Summary, Limitations and Future Work.....	72
6.1 Research Summary	72
6.2 Limitations	75
6.3 Future Work and Recommendations.....	76
References.....	77
Appendix.....	88

List of Tables

Table 1. RNN Model Error Result Before Clustering.....	44
Table 2. Result of Cluster Analysis	46
Table 3. RNN Model Clusters and Results	74

List of Figures

Figure 1. Research Methodology	4
Figure 2. Methodology of Developing the Forecasting Model.....	23
Figure 3. ERD	30
Figure 4. Organization Breakdown Structure Schema.....	31
Figure 5. Work Package Attributes.....	32
Figure 6. Project Attributes	33
Figure 7. Resource Attributes	33
Figure 8. Distribution of Work Packages Duration	37
Figure 9. Distribution of Work Package Types	38
Figure 10. Distribution of Resource Categories.....	39
Figure 11. Development of the Forecasting Model	43
Figure 12. Clustering Analysis Process in RapidMiner	46
Figure 13. Sample Illustration of Using Dummy Variables	48
Figure 14. Baseline Model Schema	53
Figure 15. Linear Model Schema.....	54
Figure 16. RNN Model Schema.....	55
Figure 17. RNN Model Structure.....	56
Figure 18. MAE of the Forecasting Model	58
Figure 19. Interface of the Application.....	62
Figure 20. Sample Run of the Application	63
Figure 21. Initial Version of the Interface.....	64
Figure 22. Main Window & Dropdown Buttons (Inputs).....	65
Figure 23. Dropdown Buttons and Lists	66
Figure 24. Main Frame Setup Code	66
Figure 25. Dropdown Buttons Setup Code	67
Figure 26. Model's Output Setup Code.....	68
Figure 27. Output Plotting Setup Code.....	68
Figure 28. Model's Prediction for Required Labor Resource in a Steel Work Package	69
Figure 29. Predicted Labor Hours Required for a Piping Work Package	70
Figure 30. Actual Labor-Hours Spent for a Piping Work Package.....	71
Figure 31. Cluster Analysis Result	73

Chapter 1 Introduction

1.1 Background

The construction industry has the highest share of employment among all industries in most countries. Also, labor costs include 30-50% of total project costs in the construction industry (Heravi and Eslamdoost 2015). Thus, effective labor resource management is essential for the success of construction projects. Through labor resource management, the demand for labor resources can be met according to the needs of a project, while achieving a realistic and accurate estimate of the demand for labor resources. However, effective labor resource estimation is challenging as construction labor productivity depends on multiple factors.

Currently, labor resource estimating is mostly carried out by project managers utilizing the common software tools and their own judgements. So, the current practices are mostly relied on the knowledge and experience of the experts involved in a project which would lead to inaccurate estimates. Moreover, with the current approach, the potential ability of the historical data collected from previous projects is ignored.

Also, with the rapid advancements in machine learning algorithms, they could be used as the powerful tools for developing forecasting models and utilizing the historical data to achieve an accurate labor resource estimation. Despite the wide adoption of machine learning algorithms in the past two decades (Gondia et al. 2020), adapting machine learning for labor resource prediction is still in its infancy (Iwu 2016; Gondia et al. 2020).

Accordingly, the goal of this study is to develop an efficient predictive data analytics model to analyze and learn from data based on previous construction projects in order to have a precise estimate of needed labor resource for different types of work packages, which helps construction

industries to utilize their collected data and to predict labor resources requirements for their planned projects precisely.

1.2 Research Objectives

This research aims to develop a machine learning model that would forecast the labor resource requirements for different work package types. The proposed objective is accomplished by achieving the following: (1) exploring the current industry practices in estimating the labor resource requirements; (2) identifying the key factors impacting the labor resource allocation in projects; (3) evaluating novel machine learning algorithms and their capacity in time series forecasting; (4) developing a generic prediction model capable to forecast the required daily labor hours for a given work package.

The academic and industrial contributions of this research are as follows:

Academic contributions:

1. Identifying the current industry practices in labor resource estimation.
2. Exploring the significant factors attributes impacting the labor resource requirements in a construction project.
3. Developing a forecasting model utilizing machine learning algorithms to predict the labor resource requirements on the work package level.

Industrial contributions:

1. Proposing a generic data acquisition model to collect project data, providing project managers a precise insight on their previous projects.
2. A framework to forecast labor resource requirements on work package level utilizing company's own previous projects.

3. Providing detailed prediction for the required labor hours instead of limiting the estimation into a total labor hour for the whole work package.

1.3 Research Methodology

To achieve the research objectives, the following tasks are performed: (1) investigating the factors that impact the labor resource requirements significantly and assessing their applicability to the defined problem; (2) exploring the current practices in the construction industry for labor resources estimating; (3) preparing a data acquisition model and collecting historical data based on it; (4) implementing feature selection to identify contributing factors and input variables of the model; (5) developing a forecasting model utilizing machine learning algorithms including linear regression and recurrent neural network; (6) training the model with the pre-processed data and evaluating the performance of the algorithms; (7) outputting the predicted labor resource hours for any new work packages by running the trained model.

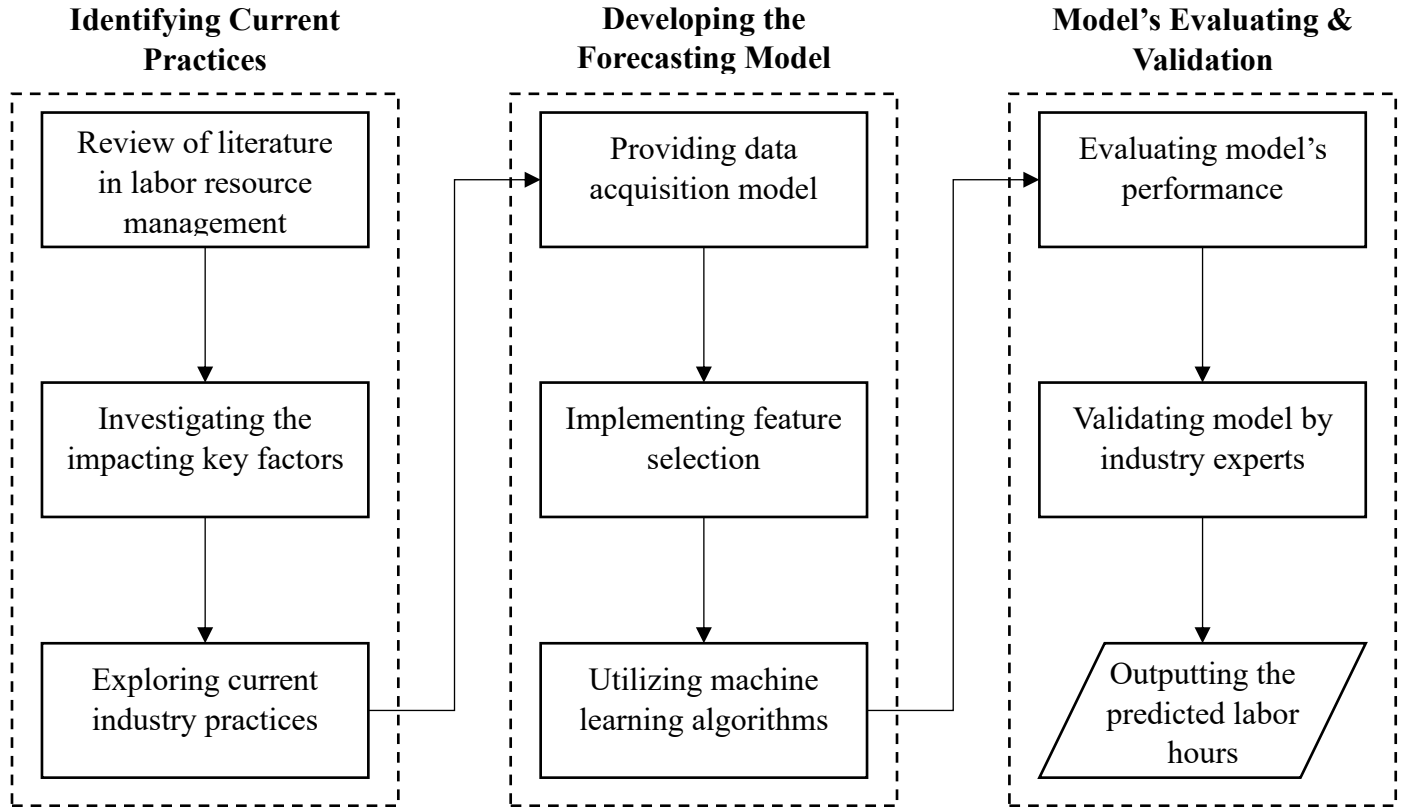


Figure 1. Research Methodology

1.4 Thesis Organization

This thesis is organized into six chapters as follows. Chapter 2 reviews the literature and novel methods in labor resource management. It also covers the use of machine learning algorithms in construction projects. Chapter 3 comprises an overview of the proposed research methodology with a detailed explanation of the data collection process and the proposed data acquisition model. Chapter 4 covers the development of the forecasting model utilizing machine learning algorithm and its implementation in a case study. Chapter 5 elaborates the developed computer application of the forecasting model. Finally, chapter 6 includes the summary of this research along with the limitations and suggested future work.

Chapter 2 Literature Review

The main objective of this thesis is development of a prediction model for accurate estimation of Human Resource (HR) in construction projects. This model is founded based on the machine learning algorithms. The review on prior work touches three different areas: Human Resource Management (HRM) in construction, factors affecting labor resource demand in construction projects and applying machine learning algorithms in construction projects.

2.1 Human Resource Management (HRM) in Construction

The construction industry has the highest share of employment among all industries in most countries (Gurmu and Ongkowijoyo 2020). Also, labor costs include 30-50% of total project costs in the construction industry (Heravi and Eslamdoost 2015). The construction industry is among the most dynamic and complex environments (Druker et al. 1996; Wild 2002; Loosemore et al. 2003) and HR-related issues is considered as a major cause of this complexity. Therefore, many theories and methods of the management of HR have been developed. How to treat, evaluate, and improve the human-related value of HR, has become an important area of research in the strategic management in construction industry (Baron 2003; Arnett et al. 2002; Pinker and Larson 2003; Wang and Yao 1999).

Accordingly, effective HRM is essential for the success of construction projects (Liu and Ballard 2008; Heravi and Eslamdoost 2015; Gurmu and Ongkowijoyo 2020). There has been a widespread realization that construction industry must improve its HRM performance before it can improve its overall efficiency, productivity, and cost effectiveness (Hammad 2009; Loosemore et al. 2003). Effective HRM practices are found to lead to positive organizational outcomes (Becker and Gerhart 1996) such as turnover (Huselid 1995) and productivity (Katz et al. 1987).

In the 1990s a gradual redefinition of the HRM from a personnel advisory role to a performance-based management activity has happened. However, a precise definition of HRM is difficult since the issue has been subject to debate within construction management experts and no one has yet provided a single, authoritative definition of what the concept means. This is mainly due to the uncertainty of whether the change in title- from personnel management to HRM- has necessarily coincided with an actual change in the way that organizations manage people or not. Consequently, there has been still discussion between leading researchers and practitioners as to what the distinction between the terms is and what this actually means in practice. Somehow, HRM can be defined as the strategies or policies and practices related to developing the HR of an organization (Inyang 2011). Due to the important role of HRs on productivity (Jantan et al. 2009) and safety (Widyanty et al. 2020), there has been an increased trend -as mentioned previously- toward HRM (Wong et al. 2004; Amrutha and Geetha 2020).

HRM includes the development of knowledge and expertise, and the enhancement of performance (Garavan and Morley 2006). A powerful HRM system is also recognized as a substantial asset for construction companies, as an organization productivity is highly correlated with its strategies (Chen et al. 2003). The development of employees, their eligibilities and the total development of the organization are the main concerns of HRM.

Researches have demonstrated the positive relationship between HRM practices and organizational performance to help companies achieve their goals (Antonioli et al. 2013; Buller and McEvoy 2012; Datta et al. 2005; Huselid 1995; Pfeffer 1998). HRM practices in construction companies could be defined as plans involved in eliminating HR-related issues in such processes including recruiting, screening, training, rewarding, and appraising the performance of HRs within companies (Bowen and Ostroff 2004; Dessler 2012; Huselid 1995). But, in the context of projects

and considering the common constraints in time and budget as well as the expectations from projects' deliverables, common HRM practices could not always be applied (Lim and Mohamed 1999; PMI 2013).

With rapid changes in technology, workers' needs, current market, financial issues and competitive environment, planning for HRs have become an important and challenging task for HRM development. HR planning involves plans for future needs of employees, their required skills, acquisition of employees, and personnel development (Werther and Davis 1982). Through Human Resource Planning (HRP), the demand for HRs can be met according to the needs of a project, while achieving a realistic and accurate estimate of the demand for HRs (Mutua 2019; Reilley 2003). However, effective HRP is challenging as it depends on multiple factors (Heravi and Eslamdoost 2015). Accordingly, some researchers have adapted Artificial Intelligence (AI) to address the existing challenges (Jantan et al. 2009). Utilizing AI in construction projects will be discussed later in this chapter.

The primary concern of HRP is to integrate the strategic and operational requirements of the project with a workforce equipped to provide the needed services and products (Marchington and Wilkinson 2002). Some research emphasizes the importance of planning, especially within the dynamic projects like construction as it can help reduce uncertainty, introduce structure and create order and action (Laufer et al. 1999). Besides, Turner (2002) represents two primary components of productive HRP: establishing a strategic HR forecast and preparing HRP framework. The strategic HR forecasting is a key input to the direction of an organization. HRP is developed to act as a means to achieve strategic HRM targets, and thus forms the output (Turner 2002).

As some research argue, planning is necessary for increasing competency levels of HRs and it causes a more efficient management of HRs in an organizational (Tsui 1987; Tabassi et al. 2012).

Additionally, planning to improve HR-related issues such as capabilities, collaboration, and team working skills could eradicate majority of HR-related risks that might appear during projects (Baiden and Price 2011; Bredin 2008; Campion et al. 1993).

Traditionally, HRP is carried out reactively during the execution of a project, instead of proactively through consideration of long-term requirements of projects and employees (Raiden et al. 2004). In construction, project managers usually prefer to allocate their expertise more to solve issues relating to scheduling, budgeting, risk management, and controlling in projects and mostly disregard HR-related issues (Scott-Young and Samson 2008; Zwikael and Unger-Aviram 2010). Besides, there are few studies which inquire influential factors that can lead to a more efficient HRM within construction projects (Belout and Gauvreau 2004; Huemann 2010). Also, the main overlooked aspect of HRP frameworks is that they are generally subsets of more comprehensive project management frameworks and are not specifically considered as an independent framework developed for the HRs (Davis 2014; Pinto and Prescott 1988). Overall, the studies which investigate influential factors that could lead to successful HRM are mostly general or limited to specific aspects such as HR empowerment or HR training (Huemann 2010; Raiden et al. 2004; Tabassi et al. 2012; Kukenberger et al. 2012; Pant and Baroudi 2008).

Accordingly, comprehensive HRP frameworks should be developed and modified to indicate main aspects of HRM in projects including assessment, training, recruitment, and development (Bourne and Walker 2005; Partington et al. 2005). Some studies tried to fill these gaps by trying new approaches for HRP framework (Baloh and Desouza 2012; Henver 2007). For example, Pournader et al. (2014) investigated a three-step design science approach consisting of rigor, relevance, and design cycles. Rigor and relevance cycles aim at proposing the initial HRP framework and conducting a qualitative exploratory study of two construction-engineering companies to establish

the validity of the framework, respectively. At the final step, the design cycle evaluates the applicability of the developed framework. The proposed framework is investigated by quantitatively testing through conducting a survey of 110 construction industry experts.

Traditionally, construction projects consist of several stakeholders including end users, promoters, project designers, project team, and work force. Hence, the wide variety of HRs involved in projects with different expectations would require their needs to be reflected on project's deliverables as well (Cleland and Ireland 2006; Davis 2014; Ballesteros Pérez et al. 2010). Thus, despite the fact that developing HRP frameworks for large organizations are common practices in the research literature (Becker and Huselid 2006; Wright and Boswell 2002), these frameworks usually do not consider HR issues as unique features of projects' environment within organizations (Belout and Gauvreau 2004; Huemann 2010).

Nevertheless, a number of attempts have been made to present HRP frameworks and to highlight the impacts that HRM have on construction projects. For instance, Tsui and Milkovich (1987) explored HRM through various aspects such as labor resource staffing, compensation, and training. Turner and Müller (2005) studied the fundamental competencies of project managers to lead projects towards success. Belout and Gauvreau (2004) compared the overall impact of HRs as personnel on the different aspects of a project for its successful implementation. More recently, Davis (2014) identified a limited number of HR-related factors that affect viewpoint of different parts involved in a project regarding project execution. There are also similar studies that point out benchmarks of project's success, which generally include HRM success factors (Verburg et al. 2012; Pinto and Slevin 1988). However, considerable amounts of these studies have seldom considered the applications of the HRM and its significant role in success of projects. Moreover, as an international project management standard, PMBOK (PMI 2013) introduces HRM in four

consecutive sections including ‘Plan Human Resource Management’, ‘Acquire Project Team’, ‘Develop Project Team’, and ‘Manage Project Team’. However, it seems PMBOK introduces HRM as a rather general framework in projects which could profoundly affect HRM practices in projects (Pournader et al. 2014).

In order to make the construction projects more manageable and approachable the projects can be broken-down into small work packages and each work package contains information related to the deliverables such as durations, resources, risks, etc. (PMI 2013). In the context of construction projects, the Work Breakdown Structure (WBS) is the tool that utilizes this technique. According to PMI, WBS is “deliverable oriented hierarchical decomposition of the work to be executed by the project team.”. WBS is used for many different purposes. Initially, it is used as planning tool to define and organize scope with deliverables. PMI defines deliverable as any unique product, result or service which needs to be produced to complete a task, phase, or project. (PMI 2013). WBS also can be used as a source for accurate scheduling and resource estimating. In this study to achieve a better prediction for resources, specifically human resources, the developed model will be applied for each work package. In other words, each construction project is initially broken-down into smaller work packages and consequently the prediction will be done for work packages. There are two types of WBS: Deliverable-Based and Phase-Based (PMI 2013). The main difference between these two approaches is the elements identified in the initial level of breakdown. In a Deliverable-Based approach the relationship between the deliverables and the scope is clearly elaborated. On the other hand, in Phase-Based approach the deliverables are demonstrated for each phase of the construction project.

2.2 Factors Affecting HR Demand in Construction Projects

Achieving an efficient HRP requires identifying the factors that have an effect on HR within construction projects. The simulated models reveal that the HR demand in construction projects is based on multiple different factors like project size and project type. The larger the size of the project is; the more HR is required for a particular project type. There is a wide variety of studies devoting to these influential factors and evaluating their impacts on HR demand in construction projects. For instance, Chan et al. (2003) showed the strong relationship between HR demand and project size in an analysis of 123 construction projects. Some of the most important factors are discussed as follows:

Type and Size of the Project

The HR demand for a construction project is closely related to the type of project as different construction projects tend to have a different product mix and fixed cost structure (Agapiou et al. 1995; Chan et al. 2002). For example, some trades such as plasterers and more technical skilled workers are closely associated with new housing work, whereas scaffolders have more employment opportunities from general repair and maintenance activities (Briscoe and Wilson 1993). The combination of skills also changes significantly when construction shifts from piling work to the construction of the superstructures. For instance, building a rural traditional house certainly requires more physical labor but less plant than a prefabricated building. Obviously, project size and type are important factors that dictate the extent to which specialized skills are practiced in the construction industry (Persad et al. 1995).

However, some researchers have identified a number of additional factors which have an impact on the HR requirements of construction projects. These include construction method, degree of

mechanization, project complexity, management attributes, expenditures on electrical, mechanical services, construction output and wage level.

Project Delivery & Construction Method

The construction method of an individual project and its delivery method impact the needed labor resources and combination of skills (Lemessany and Clapp 1978). For example, a residential block with traditional brick walls requires significantly more HR to execute than those, which were built using industrialized systems of construction concrete based. The increasing use of prefabrication, production activities off-site, and the use of other construction methods have caused a reduction in the demand for traditional craft skills like bricklaying, plastering and carpentry, but an increase in prefabricated elements erectors (Agapiou et al. 1995; Tang et al. 1990). Accordingly, the growing use of prefabricated components results in over 40% reduction in the consumption of HR in construction sites (Tam et al. 2006).

Moreover, the utilization of automation and mechanized methods also significantly influences the required labor resources (Ehrenberg and Smith 2003). In general, the more the automated equipment is utilized, the less the labor resources are required since automation tends to be labor saving (McConnell et al. 2003).

Project Complexity

Another factor affecting HR demand at projects is the complexity of the construction project (Ganesan et al. 1996). Gidado and Millar (1992) defined complexity as a significant factor impacting the labor demand on site including: technical complexity of the activities, amount of the overall and interdependencies in construction stages, project organization, site condition, and uncertainties of the work on site. For example, the design of the Bank of China in Hong Kong

supports the idea that structural design can potentially bring in resource savings. The total required steel was nearly half of a typical building with the same. The reduction on the usage of steel lead to less labor resource requirements for fixing and alignment of frames. Similarly, the design and use of modern hydrant systems decreases wiring requirements which leads to less associated labor service requirements (Fairweather 1986).

Handy (1985) introduces project size as a single factor in investigating the appropriate construction team organization. Wong et al. (2004) believes four attributes including overall technological complexity of overall project characteristics, site physical site condition, buildability level and complexity of coordination works are the important factors which might have an impact on the project HR demand.

Management Attributes

HR requirements are also affected by contractor's management skills such as planning, organizing, and controlling (Wong et al. 2003; Gould 2002). Appropriate coordination and utilization of resources especially labors on sites would lead to reduction in HR requirements (Ganesan et al. 1996). Enhanced HRM can lead to HRs saving in projects and better interfacing between different trades, such as electrical and mechanical trades. Detailed precise planning of site work could also cause reduction in labor resource requirements. For instance, in laying pipes and conduits, last-minute changes in design often result in abortive HR (Gruneberg 1997).

2.3 Applying Machine Learning Algorithms in Construction Projects

Construction projects comprises a number of different activities, which relate to and impact upon one another and are affected by various uncertainties, such as weather, geological characteristics and humans. Professional construction management is necessary to accomplish construction objectives efficiently and is essential to project success (Bush 1973). Due to uncertainties and the changing nature of the construction industry, practical construction management problems are complex and hard-to-predict (Li 1996). In fulfilment of these problems different tools have been successfully developed and applied in construction management. Naturally, humans are able to learn and can process complex problems even in the presence of uncertainties and insufficient information. Following the process of human inference offers an effective approach for solving construction management problems. Artificial intelligence (AI) relates to computer system designs that handle and attempt to resolve problems intelligently by emulating processes inside the human brain. As AI technology enhances the ability of computer programs to handle tasks for which humans are currently still better at handling, employing AI paradigms is appropriate in efforts to solve construction management problems (Haykin 2007; Tommelein et al. 1992). Various scientific and engineering fields have been paying increasing attention in recent years to different artificial intelligence (AI) paradigms.

Machine learning -a subset of AI- is considered as one of the top technologies used widely in different industries (Muizz et al. 2020). Machine learning is used for data modelling i.e., developing mathematical abstractions of data that can be used by computers to provide accurate prediction. Despite the wide adoption of machine learning algorithms in the past two decades (Gondia et al. 2020), adapting machine learning for construction management is still in its infancy (Jantan et al. 2009; Iwu 2016; Gondia et al. 2020).

In the construction industry, studies have been performed to make various predictions utilizing machine learning. For example, Tixier et al. (2016) performed random forest analysis and stochastic gradient tree boosting with 4400 validated datasets, and Gerassis et al. (2017) analyzed six years of accident data. They used a Bayesian network to predict the probability of the accident type for bank-related accidents. Amiri et al. (2016) analyzed five years of data in Iran using multiple-correspondence analysis, decision tree analysis, ensembles of decision tree, and the association rules method. Alizadeh et al. (2015) calculated the conditional probability of severe and fatal injury between the parameters of age, marital status, career, accident experience, and accident severity employing Bayesian theory. Furthermore, they tried to apply the achieved results in workers' training process in order to mitigate accidents and improve their insight about the perception of probable risks on site. Lastly, Chiang et al. (2018) conducted a cluster analysis of fatal accidents in Hong Kong. With the rapid advancement of machine learning technology, more approaches have aimed to use machine learning to address the challenges associated with HRM (Jantan et al. 2009; Heravi and Eslamdoost 2015; Xie 2020). Moreover, some studies have shown that using machine learning algorithms improves the precision of the HR-related process (Jantan et al. 2009; Wang et al. 2017).

In this section, a summary of common state-of-the-art machine learning algorithms mostly used in construction (KNN, ANN and Random Forest) is presented. Detailed mathematical descriptions of these techniques may be found in relevant references (Wauters and Vanhoucke 2017; Wang et al. 2016; Haykin 1999).

K-Nearest Neighbor (KNN)

K-Nearest Neighbor algorithm is one of the most fundamental and simple classification methods. The KNN algorithm is a simple algorithm that is based on predicting new records through

similarity measures. KNN classification was originally developed to deal with unknown reliable parametric estimates of probability density during discriminant analysis. Many studies have utilized this algorithm in different research areas as solutions by learning methods, mapping, and recognition. (Dang et al. 2005; Franco et al. 2001; Lee and Scholz 2006). The KNN algorithm is a powerful tool in dealing with classification and learning from massive datasets (Rosa et al. 2003).

In KNN, k refers to the number of neighbors included in the majority of the voting process (Gou et al. 2003). More specifically, KNN classifier needs a metric d and a positive integer k value (Kubat and Cooperson 2001). When a new input requires classification, the distance between the new data and training records is being calculated. Based on the specified threshold for the number of neighbors (which is k), k nearest records with the least distances are detected and selected. The class with more samples would be selected as the class of the new input. For example, if k is equal to three, the prediction of a new record will be calculated from the majority of votes from the three nearest neighbors.

As a result, the only parameter that needs to be tuned is the value of k . A smaller value of k can bias the model towards outliers and a larger value of k can make the modeling process computationally expensive (Gou et al. 2003). The optimal value of k resulting in the best performance could be identified through trial and error and there are also existing studies suggesting different formulas for the optimal value of k (Zhu et al. 2010; Lall and Sharma 1996). In KNN algorithm, different methods can be used for computing the distance. Euclidean distance metric is a simple and easy method for calculating distances in multidimensional input space. This method is widely used and can yield competitive results even compared to the most complex machine learning methods (Song et al. 2007; Duda et al. 2001).

Artificial Neural Network (ANN)

Networks are a smart technique for taking a complex system into simpler subsets to achieve clear understanding (Wu and Chan 2009). A set of nodes and connections between nodes are main components of networks (Muizz et al. 2020). The nodes are considered as computational units of networks and the connections establish the information flow between these nodes (Haykin 1999). In ANN method, the nodes are known as artificial neurons which is a computational model originated from the natural neurons. In the artificial neurons, inputs are multiplied by assigned weights and then calculated by a predefined activation function (Patterson 1996). ANNs are comprised of artificial neurons (Gershenson 2003). In other words, ANNs were built to process information in a manner similar to the human brain and consist of a set of interconnected input and output units where each link has an associated weight. The ANN technique enables modeling large, complex problems that involve many interrelated variables (Mourya and Gupta 2012). ANN method has a superior ability in prediction, pattern recognition, data compression, and decision-making (Chukwu and Adepoju 2012; Paliwal and Kumar 2009).

Recently, there are many different developed models of ANNs. The differences are mainly in the activation functions, the hybrid models, the accepted values, and the learning algorithms (Wu and Chan 2009). For example, recurrent neural network is one of the most state-of-art models of ANNs (Hibat-Allah et al. 2020). The recurrent neural network is commonly used in speech recognition and natural language processing, and modeling sequence data (Zaremba et al. 2014). The algorithm has the ability to memorize the sequential characteristics of the data and use patterns to forecast the probable scenario (Apaydin et al. 2020). The recurrent neural network allows previous outputs to be considered as inputs while having hidden layers (Apaydin et al. 2020).

ANN models usually require prolonged training time and interpreting the meaning of the calculated weights of the nodes can be difficult (Han et al. 2011). ANNs can be used in both classification and regression problems (Mourya and Gupta 2012).

Random Forest

The random forest algorithm is considered as one of the most precise forecasting methods (Wang et al. 2016). Random forest is an ensemble method for both classification and regression consisting of decision trees. Each tree consists of a root node that is divided into branches based on all possible outputs. The splitting is repeated for each branch until reaching a node where all instances have the same classification (Witten et al, 2011).

The main objective of the random forest algorithm is to provide a robust prediction model that has a better performance and is less sensitive to overfitting, through averaging several decision trees which might individually suffer from high variance (Bonaccorso 2018). The splitting approach in random forest is based on a medium level of randomness as opposed to a single decision tree (Probst et al. 2019). A randomly selected set of features in each tree is used to achieve the threshold that best splits the data. Thus, there are several trees that are trained, and each one would perform a different prediction (Probst et al. 2019). Moreover, the trees which have the best performance in a portion of the sample space, might provide inaccurate estimates in other portions (Pedregosa et al. 2011). Finally, prediction is made by aggregating (majority vote or averaging) the predictions of the ensemble (Wang et al. 2016).

The random forest algorithm represents many advantages. The main advantages of the Random Forest algorithm include the simplicity of the generated rules and the learning and classification process, while not being limited to numerical or categorical data (Gorunescu 2011). It runs

efficiently on massive datasets. The algorithm is also not sensitive to noise or over-fitting. It is capable to handle large number of inputs and meanwhile has few parameters compared to the other machine learning algorithm like ANN.

Due to its capabilities, there has been recently a trend towards utilizing random forest algorithm. Consequently, there are studies in construction industry in which researchers try to use this algorithm as the potential solution of various problems such as predicting occupational accidents, risk delay in construction projects, and project costs (Yaseen et al. 2020; Huang & Hsieh, 2020; Meharie and Shaik 2020)

2.4 Research Gaps

Labor resource has always been the most critical part of the construction resource pool. Effective management of resources especially labors is a complex issue which can be the guiding factor for the success or failure of any project. Although resource allocation has been widely studied by construction researchers, but project managers are still struggling with resource management difficulties. One of the key solutions are the forecasting models which can potentially provide practical answers for resource allocation problems.

The developed forecasting models are mainly focused on predicting the total labor demand for the entire project based on the general specifications of the project. Current models mostly consider the project as a single entity without dividing it into smaller packages. This approach provides limited insight for the required labor resources. Although these models can provide project managers with a general idea of the required resources, but they do not include many details about the predicted result. The point is that having a rough estimation about labor demand for the whole project without knowing the distribution of the demand over the time would not help project managers efficiently. On the other hand, there are some models which does predictions at the activity level. They assume each construction project as a group of sequential activities and forecast labor requirements for each of them. But in the early stages of a project there might not be sufficient details regarding project activities; therefore, making predictions for activities could lead to high errors and inappropriate resource allocating. Accordingly, there is a need for a predictive model which can provide estimations in a sufficient level such as work package level.

The current research aims to combine the current industry practices and literature to develop a forecasting model which is capable to predict the daily labor resource requirements for construction projects at the work package level. Having an insight for the demands at the work

package level is more efficient and also it is more practical than making rough estimations for the entire project. This approach also would not need detailed information about every planned activity during the project. The purpose is to have a generic model that provides project managers reliable results. This goal is fulfilled by: (1) proposing a data acquisition model to aid project managers in collecting the project information in a sufficient way which would lead to more accurate prediction; (2) developing a forecasting model utilizing machine learning algorithms to forecast the required labor resources over the project duration.

Some of the similar predictive models suggest linear regression methods, while others utilize non-linear approach to forecast the labor resources. This research explored different machine learning algorithms, including linear regression and neural networks, due to their capabilities in dealing with massive datasets and providing precise predictions.

Chapter 3 Methodology for Forecasting Labor Resources

3.1 Introduction

The objective of this research is to forecast the labor resources requirements for the work packages of an upcoming project utilizing historical project data and to develop a framework for collecting significant factors that would help in predicting the distribution of the required hours during the project. The research methodology utilizes the proposed approach by Ghazal & Hammad (2020) to develop a reliable forecasting model that can be used by experts, as shown in Figure 2. This framework is comprised of multiple steps, including understanding industry practices and techniques, investigating researchers' approach in forecasting required labor resources at the work package level, and exploring the key factors affecting the required hours for different labor resources. In the next step, a data acquisition model is proposed for collecting project information properly for forecasting labor resource requirements on a work package level. Then, historical data is collected according to the findings of the literature review and industry practices. After data cleaning and pre-processing, machine learning algorithms are utilized to identify the significant factors and develop a model that can forecast the distribution of the required hours for different types of labor resources. At the last step, the developed model is validated by the industry experts.

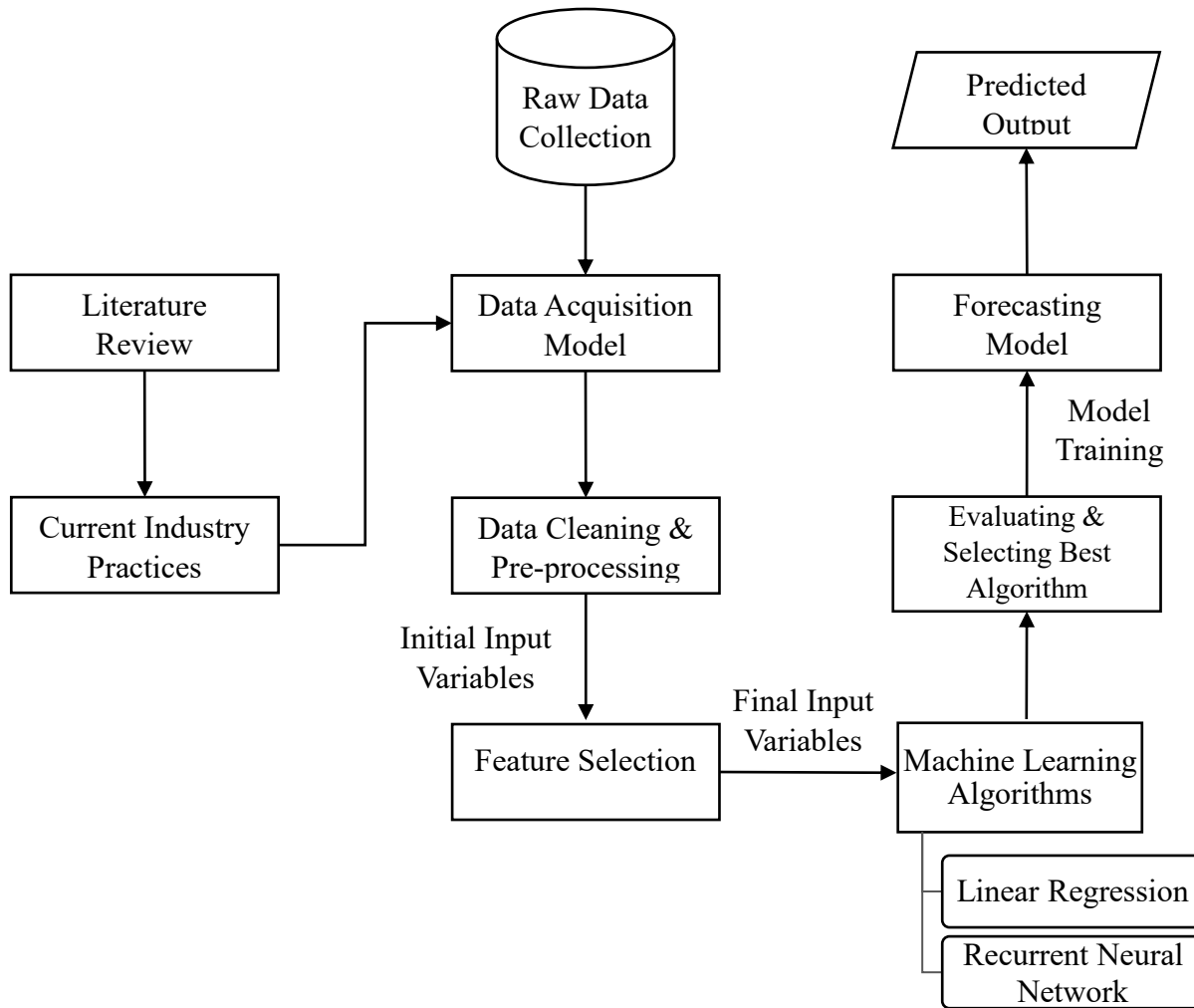


Figure 2. Methodology of Developing the Forecasting Model

The objective is to explore the project features that affect the labor requirements for different work packages of a construction project. Besides, the developed predictive model is capable to find the weights of these factors and interpret the results. It is noteworthy that the result needs to be reasonable and comprehensible to construction experts especially project managers, so it can provide them reliable support in making resource allocation and workforce planning.

The focus in this chapter is on: (1) investigating resource allocation methodologies adopted by industry practitioners; (2) identifying the significant project factors that have an influential impact

on predicting resource requirements through analyzing the previous research and discussions with experienced project managers; (3) developing a data acquisition model to ease proper data collecting for project managers; (4) collecting historical project data to be utilized in the development of the forecasting models; (5) data cleaning and pre-processing to prepare it for further exploration, analysis, and data mining; (6) describing the factors in the dataset and representing the collected data through graphs and charts to elaborate the distribution of factors.

In the next chapter, the process of forecasting model development is discussed in detail, including the data preparation approach, feature selection, defining the model inputs, and machine learning algorithms used to predict the hours required for labor resources at work package level for an upcoming construction project.

3.2 Significant Key Factors Affecting Required Labor Resources

Estimating the required workforce has always been one of the project managers' concerns. There are different methods adopted by the industry practitioners and researchers to estimate the labor resources required for a project. One of the objectives of this research is to explore the potential factors impacting the resource requirements on a work package level and to determine the significant factors that can be used to forecast labor resource requirements. Additionally, further exploration is performed to outline the current industry practices. The investigation is conducted to identify the project characteristics that need to be considered for estimation of the required labor hours for different work packages of a construction project.

3.2.1 Industry Practices in Labor Resource Estimating

Exploring the current industry practices is a critical step in detecting the key attributes required for prediction. The target is to implement an investigation approach for impacting factors considered by industry experts to estimate the labor resources required to complete a work package. Informative discussions with experienced project managers working in the largest construction companies lead to a clear understanding about the current resource allocation methodologies in the industry. Typically, each organization has its own way of identifying the required staff. They adopt 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.

Some common resource estimation methods used by industry organizations include:

1. Expert judgment: Construction organizations usually rely highly on their experts' evaluations. Companies bring in experienced specialist who have done this sort of work previously and getting their opinions on what resources are needed.
2. Project management software: The developed software such as Microsoft Project often have features designed to aid project managers to estimate resource requirements and constraints and find the best combination of assignments for the upcoming project.
3. Bottom-up estimating: Breaking down complex tasks into smaller activities and working out the resource estimating for each activity. Actually, it is the process of estimating individual activity resource need and then adding them up together to a total estimate. This method is very common among construction companies due to its accuracy and simplicity. However, it takes a considerable amount of time to perform bottom-up estimating as every activity must be assessed and estimated precisely to be included in the calculation.

4. Published estimating data: Many project managers in construction industry utilize such data to figure out the resource requirements. They rely on articles, books, or journals for analysis.

In the process of labor resource allocation, various techniques including qualitative or quantitative are adopted. However, each of these frameworks have limitations that impact the accuracy and lead the practitioners to inaccurate estimation and refrain from using their methodologies. The main drawbacks can be summarized as follows: solely relying on project managers' knowledge and experience, not analyzing the actual project values compared to the estimated values and the allocation methodologies cannot be generalized across departments and project types.

3.2.2 Selection of Factors

The key attributes impacting labor resource demand are identified by exploring the literature review and determining the factors applied by industry experts through understanding their methods in estimating the required workforce. One of the objectives of the current research is to collect the significant key features from historical data to forecast the labor hours of different resources required for work packages of a given construction project.

As mentioned in the literature review chapter, there are a few prediction models developed by researchers to estimate the labor required for a construction project. These models consider various factors in their developed forecasting models such as the project type, cost, and floor area while ignored other factors such as complexity and project delivery method (Elkholosy 2020). The current models are mostly able to do predictions only for the “total” labor hours required for a project while this research tries to develop a model which is capable to predict the required labor hours for each time step (day or week) during the project. The most frequent factors impacting the project's labor resource requirements considered by the researchers in developing their models

include: (1) project type (Yang & Kim, 2019; Chen et al., 2008; Bell & Brandenburg, 2003); (2) cost (Wong et al., 2008; Chen et al., 2008; Bell & Brandenburg, 2003); (3) work type (Yang & Kim, 2019; Othman et al., 2011); (4) floor area (Chen et al., 2008); (5) project duration (Chen et al., 2008); (6) site condition (Yang & Kim, 2019).

Additionally, some factors that are mentioned by researchers cannot be generalized to all types of projects. For example, the floor area can only be considered for estimating buildings projects.

After investigations in the literature review and discussions with industry experts the following factors were selected to be explored in this study:

1. Work package type: which describes type of the work defined for the work package such as welding, concrete, piping, cut & fill, etc.
2. Resource category: which indicates the type of labor resource used for the work package such as carpenter, ironworker, pipefitter, etc.
3. Duration: which elaborates the total time (day or week) in which a specific type of labor resource is involved in a work package.
4. Key quantity: representing the total amount of a unit of measurement to complete a work package which can be in different units depending on work package type. For example, 500 cubic meters for a concrete work package or 100 square meters for a piping work package.
5. The cost: which indicates the total dollar value spent to complete a given work package.
6. The complexity: which elaborates the technical and technological complexity of completing a work package.
7. Project phase/sub-phase: which represents the stage of the project such as planning, procurement, construction, etc.

Some of factors that might have an impact on the labor resource requirements are not considered in the case study of this research due to unavailability or confidentiality. For example, considering the impact of budget on the required labor hours is disregarded since the industry partner did not provide this information as it was considered confidential. Moreover, the complexity of a work package, which could be a factor affecting the needed labor hours, is difficult to collect as contractors do not include such information in their records and each contractor might have its own definition of complexity.

Thus, a total of 6 factors was used as an input to the forecasting models in this research. These factors include: work package type, resource category, duration, key quantity, work package cost, and complexity. The selected factors are then investigated by experienced project managers to ensure that these factors are correctly chosen. Also, the project managers are asked to provide feedback if there are any overlooked factors that need to be considered.

3.3 Data Acquisition Model for Data Collection

Appropriate data collecting is a vital and critical step for implementing the information and analyzing it. Utilizing the historical information of previous projects can improve the overall performance of the organization by increasing the efficiency and reducing the expenses (Elkholosy 2020). This will help the organizations 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.

Since most of the construction companies have their own tracking systems, each company collects the data in a different way. For example, contractors mostly focus on tracking labor hours without

considering other aspects of the projects. On the other hand, owners usually tracking the progress of the tasks and the costs rather than the spent labor hours. This variance in the process of data collection would cause difficulties in gathering datasets from various companies and implementing them into a generic model. Therefore, this study tries to propose a generic data acquisition model to provide a tracking system which is enabled to collect data from multiple projects. Then a general forecasting model is developed which can also be utilized by different construction companies with various tracking systems.

By exploring the business process of the construction contractor in the case study, it is noticed that their tracking systems is well organized and accurate due to the direct relation between the spent hours and expenses. Therefore, missing information is not a big concern while the data is pulled from their tracking systems. As the other contractors might not collected their data in such efficient way, a data acquisition model is prepared to aid in collecting the required project data for estimating the labor resources in a structured way. The model will improve the data collection and the resource allocation processes in an organization.

The provided approach facilitates the process of the storing data and allows the organizations to benefit more from their stored data and to track their project data properly in a generic way. The outcome will support the companies in utilizing the labor resources forecasting model by collecting high quality data. Also, it will help them in performing future analysis that could potentially provide them with precise insight to take corrective actions and make proactive decisions. An Entity Relationship Diagram (ERD) is developed to illustrate the different entities involved in the framework and the relationships between these elements.

3.3.1 Entity Relationship Diagram (ERD)

The initial step to achieve a well-organized acquisition model is to have a good understanding of the attributes and their connections. The goal of developing the ERD, illustrated in Figure 3, is to represent the relationships between the different entities that need to be tracked and stored in the database. The purpose is to track the labor hours spent by resources for each project work package since the required resources varies from one work package to another. Also, the quantity of resources alternates during the project, so it is far from reality to consider the hours required for a project is spent equally during the days of the project. The developed framework helps the contractors in collecting the required project data for labor resource forecasting.

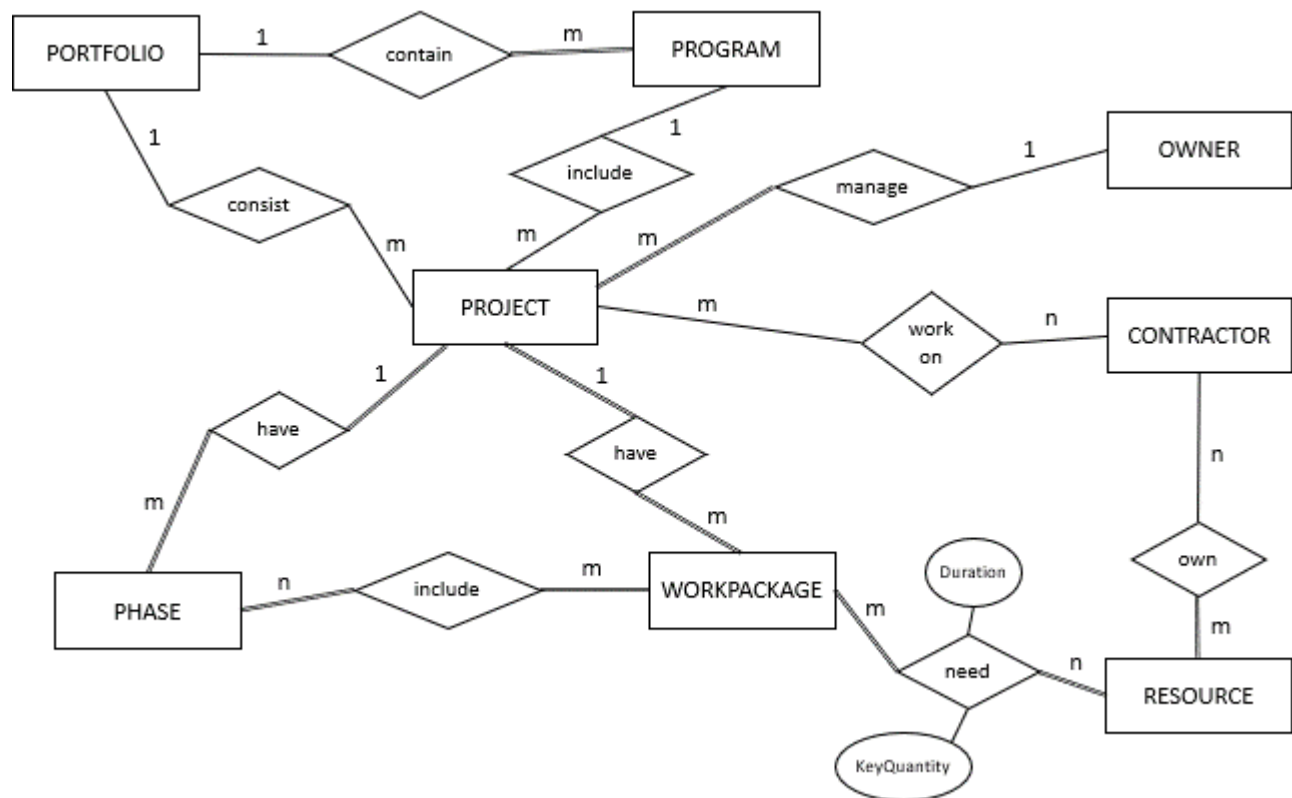


Figure 3. Entity Relationship Diagram

The framework is designed for the contractors to collect the project data in a proper way; hence, a contractor entity is included in the ERD that the other entities are built on. The designed entities

in the ERD include multiple attributes that is represented in the following section. Each organization might have multiple business units or divisions which include various departments (Figure 4). The model provides three levels of the breakdown structure to reach the level at which the projects are managed. The department have specific types of resources with the quantities that indicate the available workforce. The resource entity involves different labor types such as welder, carpenter, pipefitter, project manager, etc.

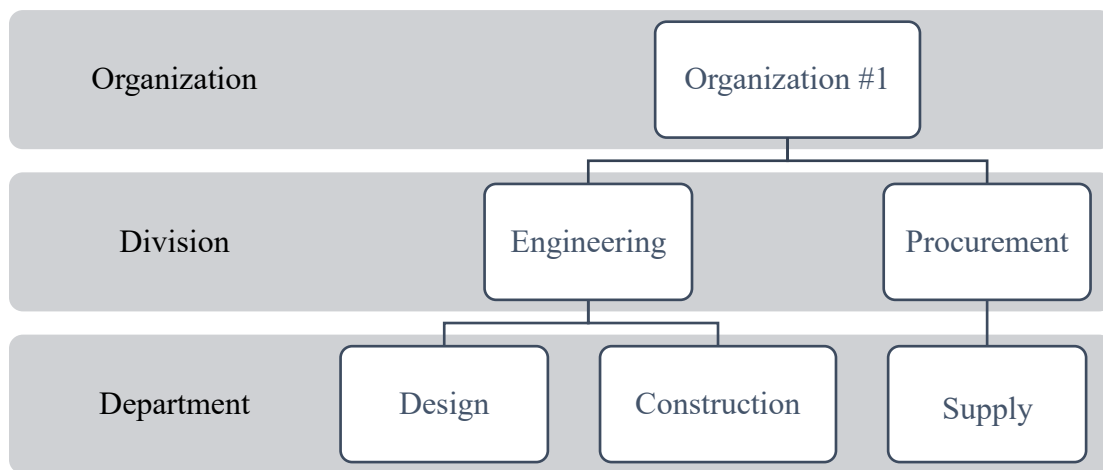


Figure 4. Organization Breakdown Structure Schema

According to the breakdown structure, a portfolio is comprised of multiple programs that would help fulfil organizational achievements through strategic business objectives along the way (Martinsuo & Lehtonen 2007). A group of related projects create the program. the program's goals cannot be completed without combining these projects (Blichfeldt & Eskerod 2008). Portfolio management is necessary for the success of the organization. It allows the company to assess, prioritize, and choose projects compatible with the overall strategy and assign resources to projects evaluating the priorities and policies (Elkholosy 2020; Meskendahl 2010).

3.3.2 Attributes Required for Forecasting Labor Resources

Each project is comprised of multiple work packages. Moreover, there are one or more contracts in a project that provide various services, including consulting, construction, soil investigation, etc. Each work package requires specific types of resources to execute its scope and deliver its requirements. Therefore, the forecasting of labor resource requirements should be performed on the work package level when the organizations store the data.

Accurate labor resource prediction cannot be conducted for each work package without considering work package and resource features. Accordingly, the project and required resources are the essential entities in this research that need investigation since the forecasting of the workforce requirements is reliant on their features. The work package and resource characteristics are illustrated in the ERD with the attributes that need to be collected in Figure 5, 6 & 7. Besides to the forecasting model, the collected data can provide the contractors more benefits such as further analysis for the inherent trends and making decisions.

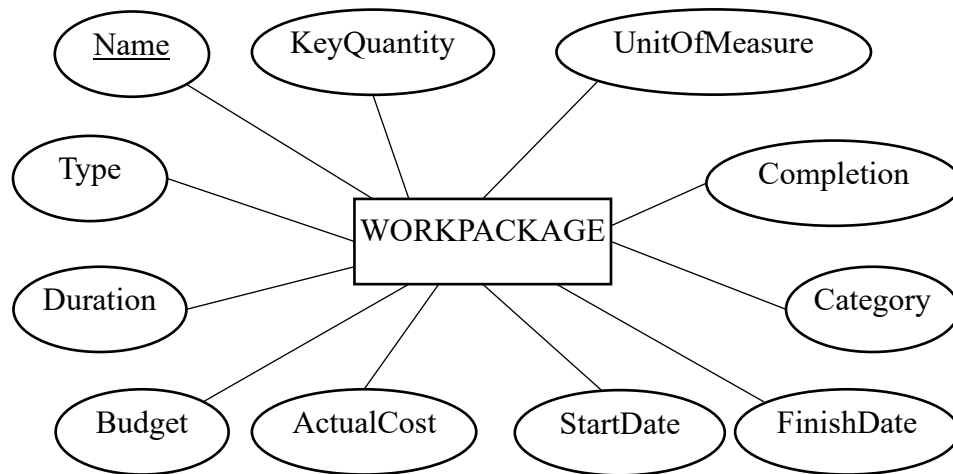


Figure 5. Work Package Attributes

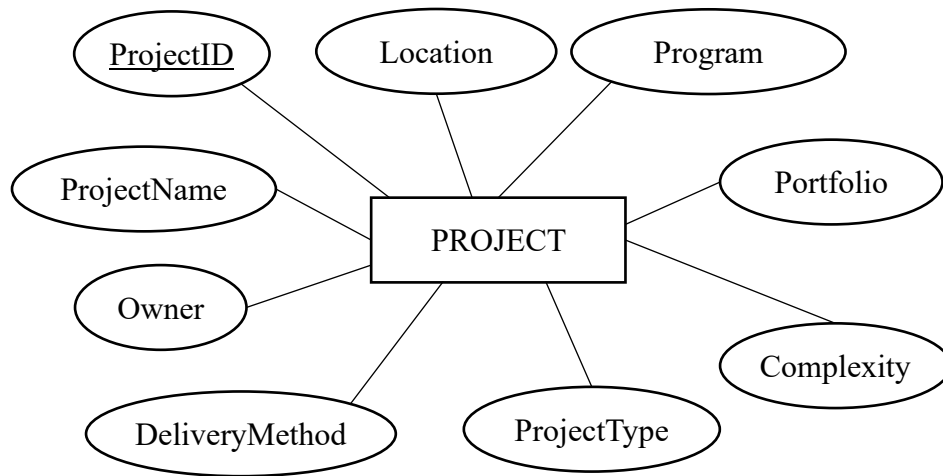


Figure 6. Project Attributes

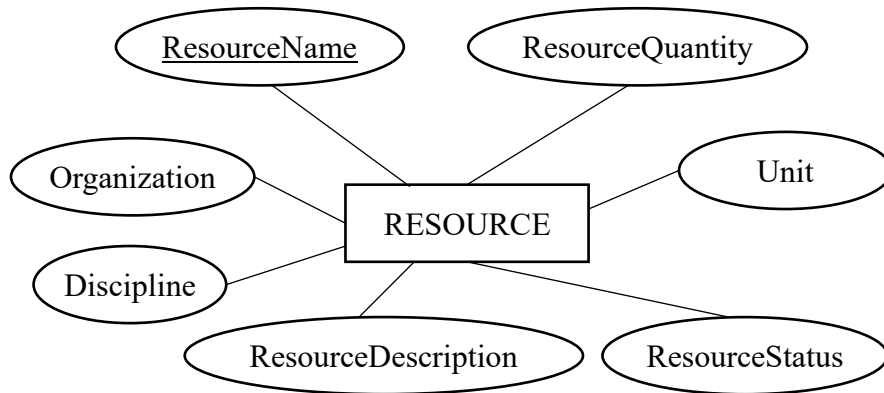


Figure 7. Resource Attributes

3.4 Data Collection Process

The high quality data for training plays an important role in the accuracy and reliability of the developed forecasting model. In this study, historical information consisting of different work packages is collected from contractors who were involved in huge construction projects taking place in 2019 and 2020.

Collecting the data is the initial step to utilize it in the forecasting models to estimate labor resource requirements. The information that needs to be collected is proposed in the literature

review and the industry practices. In this research, the purpose is to gather historical data on the work package level. In this study, the historical data is gathered through the spreadsheets, received from the industry partner, which are originally derived from company's Enterprise Resource Planning (ERP) systems.

3.4.1 Essential Attributes for the Forecasting Model

The critical factors that affect the resource demand are determined through reviewing the literature and tracking the industrial practices of construction contractors. The followings are the essential attributes of a project and its work packages: work package duration, and complexity; work package type (concrete, excavation, steel, etc.), project cost, resource category (carpenter, pipefitter, ironworker, etc.), and key quantity. In this study, the data collection is focused on collecting the attributes for the historical projects which are completed. The labor resource hours spent on the work packages are collected and used to train the forecasting model.

3.4.2 Data Sources

Each contractor has its own way of collecting project data. Some might use their own developed collecting systems, or some may utilize the ERP systems such as Oracle.

The industry partner involved in this study has developed its own tracking system. The tracking system includes different spreadsheets each used for collecting specific data. The hours spent on each task is collected on a daily basis. Each task has a unique function code which is the key to link spreadsheets to each other. These report spreadsheets include the task description, the break down structure for activities, function description, employee type, work date, planned key quantity, actual key quantity, and hours spent on each day.

3.4.3 Missing Information

Although the contractors usually track the tasks and their spent hours accurately, but there are some attributes which contractors mostly do not collect such as complexity or location. Moreover, they might not use the same proper format for collecting data from all projects. For example, without having a well-defined format, naming different tasks or resources can get complicated and confusing as each project manager might use his/her own way of naming various tasks or resources. Having a predefined data acquisition system can prevent all these problems and provides a clean and well-organized dataset which can really benefit the company in accurate analysis, quick spotting any possible issue and future planning.

The industry partner involved in this study has a powerful tracking system which is developed within the company and has been utilized and improved during years. As a result, there were only limited missing values in their collected data. But as mentioned before, there are some attributes which the contractor usually does not include in their tracking systems. The missing values are removed from the training dataset through outlier detection process which will be discussed in detail later.

3.4.4 Data Transformation and Linking

As there might be different timesheets for collecting various aspects of the project and most of these attributes may not be necessary for the research purpose, developing a well-defined database is required. This would help to collect the needed data in a structured manner. Hence, the collected data from different systems is collected in an organized database ready for importing into the forecasting models.

The historical data collected from the contractor was compromised of different spreadsheets. Each of these spreadsheets covered a specific aspect of the project. For example, one was the timesheet

of hours spent on each day describing the tasks and resources used on it. The contractor has a predefined hierarchical list of tasks and resources involved in a project calling each level “parental code”. There are 3 levels of “parental code” defining the break down structure of each task. For example, “Deepen Ditch along West Perimeter Road” is a task done during the project and is the lowest break down level called “parental code 3”. The upper level of this task “parental code 2” is defined as “Site Drainage” and the highest level “parental code 1” is called “Sirte Drainage WO1 – Athabasca”. Another spreadsheet included the actual and planned key quantities for the project tasks. Each task has a unique “functional code” which act as the key attribute for linking different spreadsheets together.

There were various labor resources who worked on the work packages. Therefore, it was essential to aggregate the similar types of resources as one resource type, because then there would be too many resource types and the forecasting models were not able to perform an acceptable prediction.

3.4.5 Data Understanding and Visualization

6 key attributes which significantly impact labor resource requirements are identified. They are used as inputs into the forecasting models. The following section briefly explains these factors with the illustrations of their distribution in the collected dataset.

3.4.6.1 Work Package Duration

The work package duration is in days, and it represents the time between start and finish date represented in the tracking system. The purpose of applying this factor is to indicate the effect of the duration on the labor resource hours required to finish the work package.

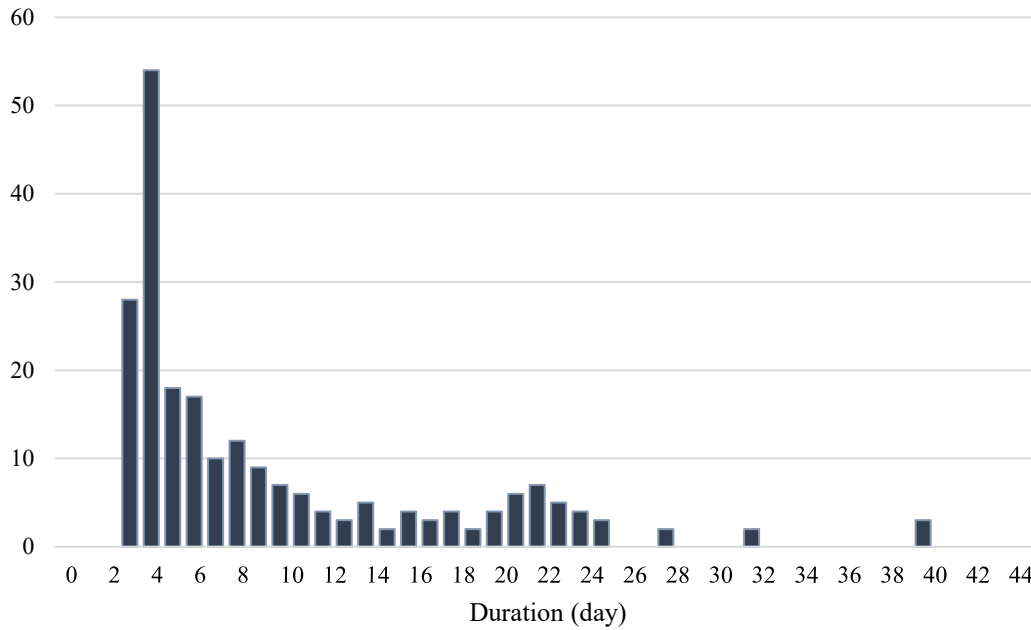


Figure 8. Distribution of Work Packages Duration

3.4.6.2 Work Package Type

Work package type represents the category of the work that is done such as bolt up, cut & fill, or piping. In the case study there are 11 types of work packages including: Concrete, Culvert, Cut & Fill, Excavation, General, Piping, Rework, Services, Site Drainage, Steel, and Superintend.

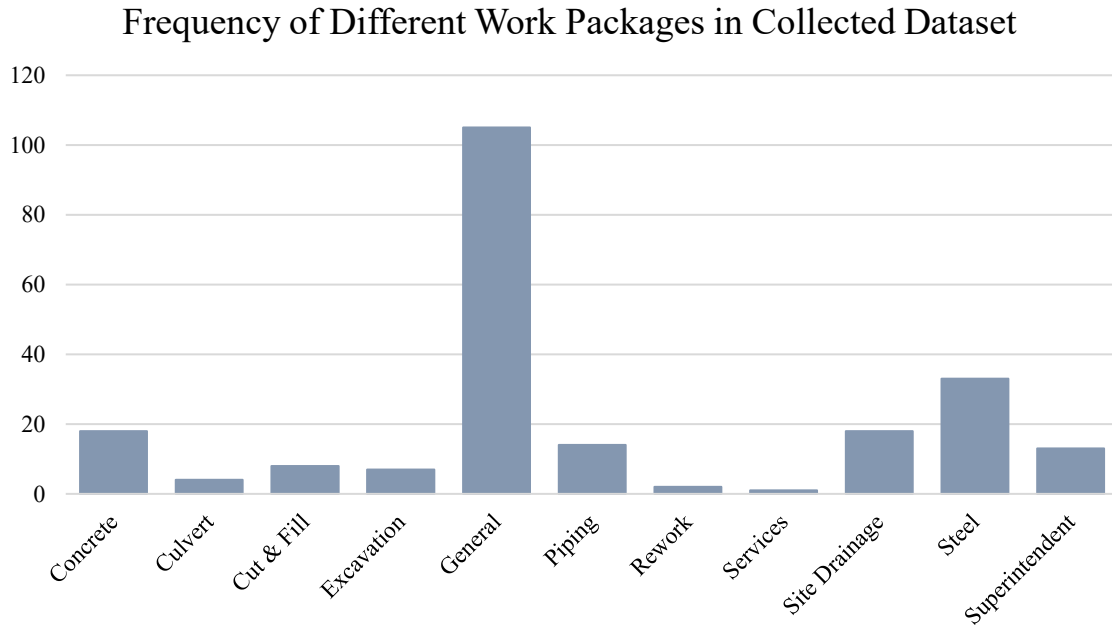


Figure 9. Distribution of Work Package Types

3.4.6.3 Work Package Complexity

The complexity of the work package is defined through a 5-level scale to represent the technical complexities and the uncertainties. The scale is designed to measure the required effort for completion of work packages with different level of complexities. As mentioned before, there are some attributed which the contractors do not include in their tracking systems or there might not be defined way to measure such ones. The complexity is among those attributes which is not included in their dataset. Accordingly, this attribute is not included in the case study due to unavailability.

3.4.6.4 Work Package Cost

The total work package cost represents the value spent to complete the work package. The higher the total cost of the project, the bigger the work package size is. However, due to confidentiality of such information the contractor did not include the costs in the given spreadsheets.

3.4.6.5 Resource Category

The resource category represents the type of labor resource that are utilized in the work package. Each category includes all the different subcategories of the specified resource. For example, construction estimator, construction inspector and senior project manager are different labor types, but they are all categorized in one type as project manager. The labor resource categories in the case study include 12 types: Project Management, Overburden Operator, Carpenter, Welder, HSE, Ironworker, Labor, Operator, Pipefitter, Project Execution, Quality, and Accounting.

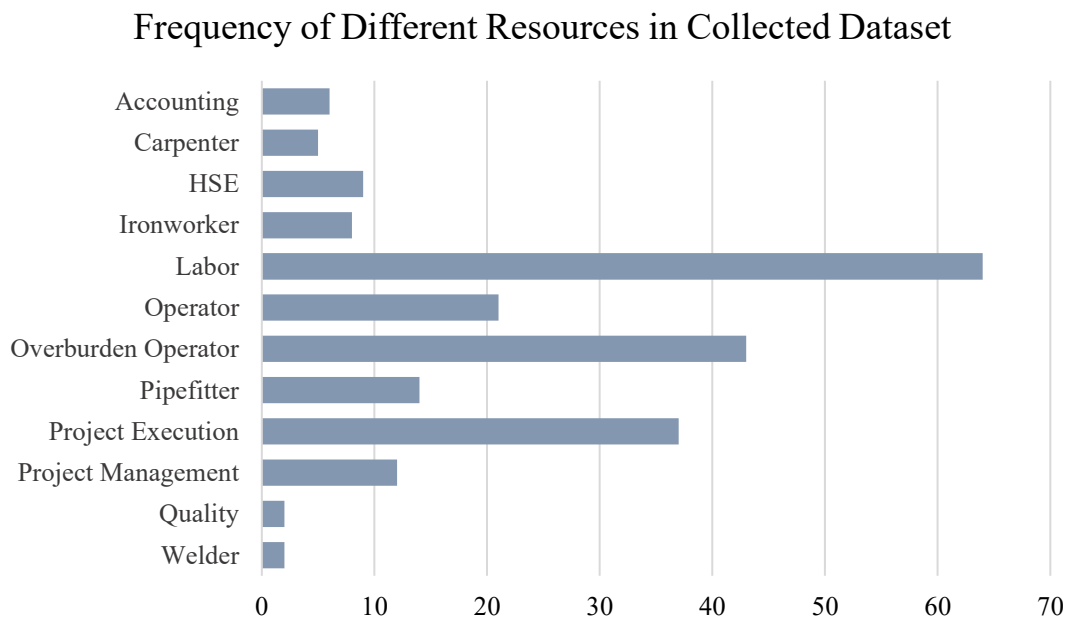


Figure 10. Distribution of Resource Categories

3.4.6.6 Key Quantity

The key quantity refers to the amount of main material, service, etc. that represents the scope of a work package. Based on the work package type the unit of measure could be various. For instance, the key quantity of a concrete work package is measured in cubic meters. The key quantity for

pipings are measured in inch/diameter. As a result, comparing key quantities is only acceptable when they have the same unit of measurements.

3.4.6 Deficiencies in the Collected Dataset

There are a few limitations that were encountered during the data collection process. These limitations are mostly due to the confidentiality or unavailability of some information from the contractor side. Therefore, some assumptions were made to have a target dataset to forward in this research. These limitations and assumptions include the following:

(1) The actual duration of each work package is calculated by subtracting the first date included in the spreadsheet from the latest date which is not the accurate duration of the work package. In some cases, there are several days with zero working hours in the first or last days of a work package which is not clear whether the working hours are missed on those days, or they should not be included in the duration. Therefore, it is assumed that there were no works on those days, and they are excluded from the work package duration.

(2) The actual cost and budget were not available due to confidential issues. As a result, the case study was done without cost attribute. Considering cost as an input would potentially improve the performance of the forecasting model. However, the total number of working hours in each package is available in the dataset.

(3) In calculating the duration, the holidays and weekends were excluded from the duration of the work packages.

(4) In general work packages which include recurrent tasks such as maintenance and safety, the durations were significantly large (near to project's duration); so, in order to avoid high error

values in the models, these types of work packages were limited into a few records which had the acceptable durations.

(5) The complexity of the work packages was not defined in the spreadsheets; and as there is not a standard method for measuring the complexity, the complexity was also excluded from input variables of the forecasting models.

(6) The data that is used for case study is limited into the projects which their work packages are mostly limited into some specific types such as piping. Accordingly, this limitation does not allow to train the model with different work packages of projects with different characteristics.

Chapter 4 Development of Forecasting Model for Labor Resources

4.1 Introduction

As previously mentioned, the objective of this study is to identify significant factors impacting labor resource requirements using feature selection techniques, to develop a machine learning model to predict the labor resources required for construction work packages to help project managers in workforce allocation during the early stages of planning.

According to the applied methodology in the previous chapter, 6 factors affecting the labor resource requirements were identified. In this chapter, a framework is proposed for development of a forecasting model for labor resources where the identified factors are used as inputs.

To fulfil the objective of the research, an accurate and reliable forecasting models needs to be developed by utilizing machine learning algorithms. Figure 11 indicates the steps of the following framework to build the forecasting model. At the first step, the collected dataset is cleaned and processed. Then, the clustering analysis is implemented to detect the optimum number of clusters and their ranges. Also, the anomalies are identified in the dataset through outlier detection techniques. Then, the feature selection methods are applied to investigate the feature with the huge impacts on the performance of the prediction model. At the final step, different machine learning algorithms are explored through their capabilities to select the most suitable algorithm for the purpose of this study. The chosen algorithm is utilized in the development of the forecasting model. Then, performance of the developed prediction model is evaluated by measuring the error value.

In this chapter, first data preprocessing and cluster analysis are explained. Then, feature selection approach is elaborated. After that the selected machine learning algorithm and the developed

forecasting model are discussed. And finally, the performance evaluation of the forecasting model and validation process are reviewed in detail.

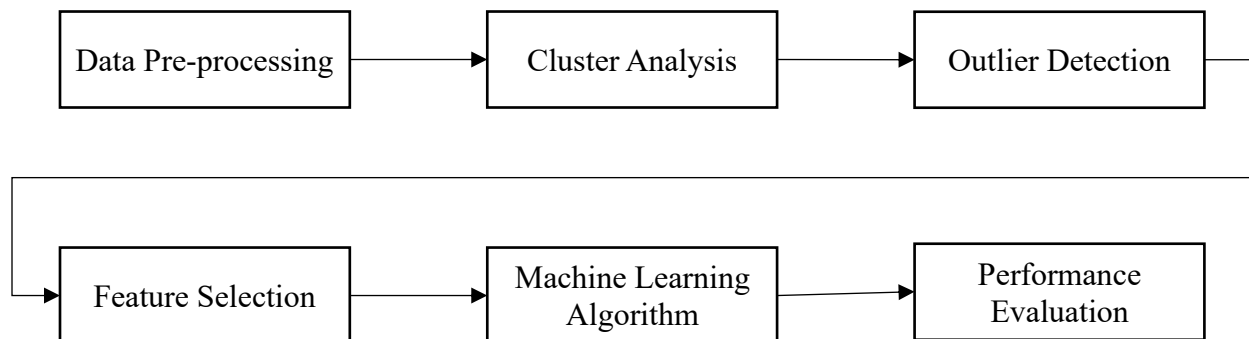


Figure 11. Development of the Forecasting Model

4.2 Data Preprocessing

Most of the machine learning algorithms require some data cleaning and preprocessing prior to any training. The performance of algorithms highly depends on the input data. So, the data needs to be prepared correctly compatible with the algorithm to prevent any misleading or inaccurate results. One of the steps in data preparation is dealing with the missing values found in the dataset. The missing information is mainly caused by improper tracking during the project. Also, outliers can be found in the stored datasets where have been deficiencies in the tracking and storing process. At the next step, the attributes in the dataset need to be scaled to prevent the forecasting model from assigning unacceptable weights to the attributes because of greater values. Moreover, the categorical attributes be set up. Handling the categorical attributes is another critical and necessary part of preprocessing since some of the algorithms only take numerical inputs such as regression and neural networks. Accordingly, the required modifications should be applied on the dataset before proceeding into the training process.

So, proper formatting is necessary for the input dataset due to the discussed reasons. This process includes some essential steps, in order to achieve legitimate input data for the model, which are removing the records with missing values, eliminating outliers, handling categorical project attributes, and normalizing the training dataset.

4.2.1. Missing Values

In the collected data some records did not include the hours spent on several days or the start and end dates were not clear for some work packages. This occurred due to lack of allocation of hours to projects on timesheets or the absence of timesheets to track the spent hours on some days. Accordingly, some of the work packages were removed from the input data as it was unfeasible to collect this information in any other way.

4.2.2. Cluster Analysis and Outliers Detection

As the duration of the work packages varies from a few days to several weeks, so it can obviously be concluded that the model would not perform accurate prediction due to this huge variety. Hence, a cluster analysis can improve the performance considerably. Table 1 indicates the error value for the model before applying cluster analysis. As shown in the table, the Mean Absolute Error (MAE) of the model is considerably high. Therefore, clustering and grouping the relevant work packages is necessary for the performance enhancement. The details of the MAE and its calculations are discussed later in Section 4.5.

Table 1. RNN Model Error Result Before Clustering

Cluster	Number of Work Packages	Duration (day)	MAE (labor-hours)
1	223	>2	± 14209

Cluster analysis is a statistical method for processing data. It works by organizing items into groups, or clusters, on the basis of how closely associated they are. Cluster analysis is an unsupervised learning algorithm, meaning that you do not know how many clusters exist in the data before running the model. Unlike many other statistical methods, cluster analysis is typically used when there is no assumption made about the likely relationships within the data. It provides information about where associations and patterns in data exist, but not what those might be or what they mean. Accordingly, it is a powerful data mining tool for identifying discrete groups of work packages based on their durations.

k-means is commonly used algorithm in cluster analysis. k-means algorithm establishes the presence of clusters by finding their centroid points. A centroid point is the average of all the data points in the cluster. By iteratively assessing the Euclidean distance (Equation 1) between each point in the dataset, each one can be assigned to a cluster. The centroid points are random to begin with and will change each time as the process is carried out. Figure 12 indicates the process of cluster analysis and outlier detection in the RapidMiner software. The result suggested that the optimum number of clusters is 4. The boundary of each cluster is shown in the Table 2. It is noteworthy that the outlier detection process is simultaneously applied with the clustering process. The outliers are automatically removed from the data set.

$$d(p, q) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2} \quad \text{where } p, q = \text{two point in Euclidean } n - \text{space} \quad (1)$$

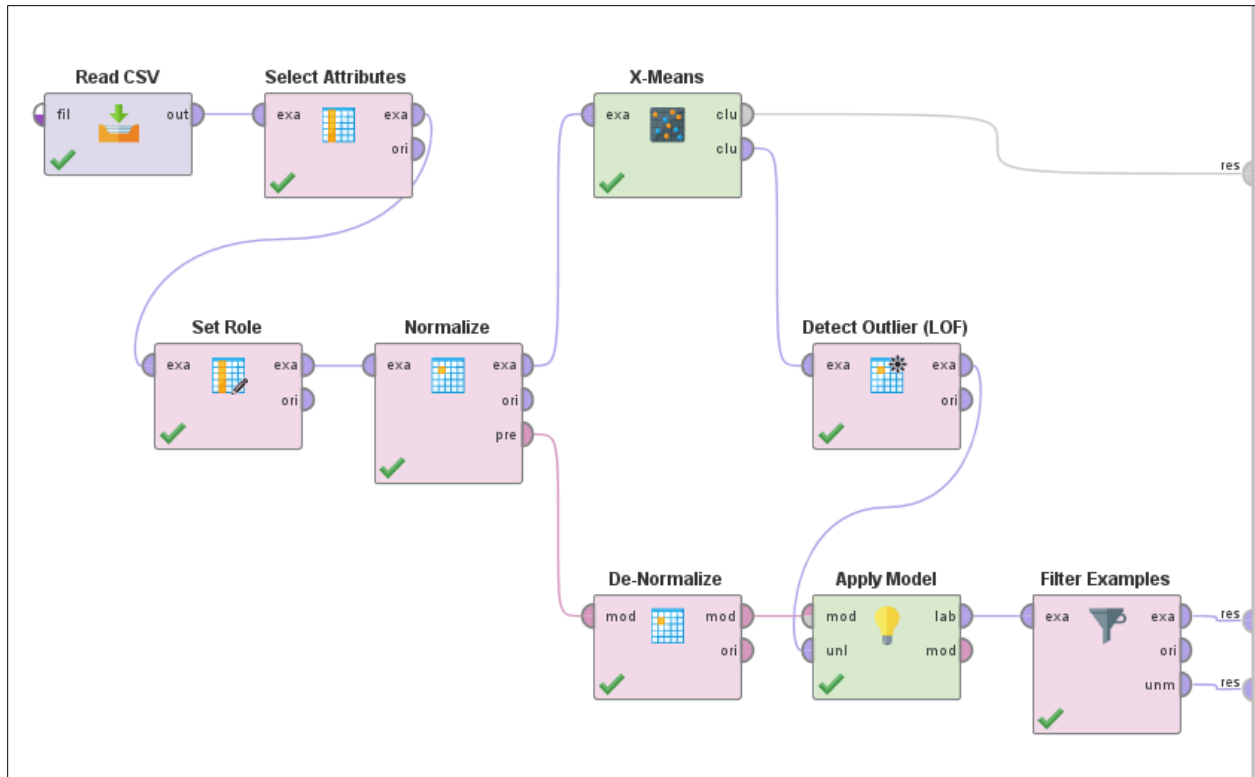


Figure 12. Clustering Analysis Process in RapidMiner

Table 2. Result of Cluster Analysis

Cluster No.	Number of Work Packages	Duration (day)	
		Lower Bound	Upper Bound
1	7	26	-
2	37	16	25
3	34	9	15
4	145	0	8

The following is a brief description of how clustering is done:

1. First, the data is prepared for the analysis. Normalizing of the data is a typical step when comparing attributes of different characteristics. In the proposed approach z-transformation method is used to ensure that deviations are equal, so that the outlier has a clear meaning.
2. The clustering operator is applied to the data to find coherent groups in the dataset. Then, the outliers are explored using the local outlier factor mechanism which identifies outliers based on local outlier factors where locality is given by the k nearest neighbors. By comparing the local density of a record and the density of its neighbors, the record is considered as an outlier if its density is lower than the density of its neighbors.
3. The data is de-normalized by applying the reverse normalization model, thereby obtaining the original dataset. Then, the examples are filtered to get one dataset with the outliers and another with the rest.

4.2.3. Nominal Features

Dealing with categorical attributes might not be as easy as numerical ones. So, there are some techniques to transform them into numerical variables. This transforming is an essential process of data preparation. The prepared data set included two types of categorical data: nominal attributes, such as work package type and resource category, and ordinal attribute like complexity. Nominal ones 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.

One of the techniques to deal with these variables is to use dummy variables. Hence, dummy coding is 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 instance, new columns will be created for each work package type and the type will be specified

with a value of 1; the remaining columns will be 0 (Figure 13). All categories within each categorical attribute are assigned a dummy value. This technique 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 (Elkholosy 2020).

Work Package Type		WP-Steel	WP-Concrete	WP-Piping	WP-Cut & Fill	WP-Culvert
Steel		1	0	0	0	0
Concrete		0	1	0	0	0
Piping		0	0	1	0	0
Cut & Fill		0	0	0	1	0
Culvert		0	0	0	0	1

Figure 13. Sample Illustration of Using Dummy Variables

Moreover, the work package complexity involved ordinal attributes with values of very low, low, medium, high, very high. Therefore, assigning the proper numerical values to the 5-rating scale is suitable for a good forecasting model. A value from 1 to 5 is assigned to the different levels of the complexity, so the algorithms could recognize work packages with higher complexity should have more weight allocated to them.

4.2.4. Normalizing & Data Splitting

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. It is important to scale features before training a machine learning model. Normalization is a common way of doing this scaling. Subtract the mean and divide by the standard deviation of each feature.

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. The mean and standard deviation should only be computed using the training data so that the models have no access to the values in the validation and test sets. It is also arguable that the model should not have access to future values in the training set when training, and that this normalization should be done using moving averages.

Then, to provide an unbiased estimate of learning performance, the training data set is randomly divided into 3 different subsets: (1) Training set is a subset of the dataset used to build predictive models. In the developed model training subset consists of 70% of the records. (2) Validating set is a subset of the dataset used to assess the performance of the model built in the training phase. It provides a test platform for fine-tuning a model's parameters and selecting the best performing model. Validation set includes 20% of data set in the current model. (3) Testing set, or unseen data, is a subset of the dataset used to assess the likely future performance of a model. If a model fits to the training set much better than it fits the test set, overfitting is probably the cause. In the current model 10% of data set is dedicated to the test subset.

The k-fold ($k=8$) cross validation technique is adopted for training and validation of the models. As a result, the data records are randomly divided into 8 equal clusters and the train-validation process is repeated for 8 runs, where each cluster is used for validation and the remaining clusters are used to train the models. This process is repeated until all clusters are selected as the validation set. This ensures that the dataset is completely utilized as the training set and leads to improvement in the predictive performance of the model (Poh et al. 2018).

4.3 Feature Selection

Feature selection techniques are frequently used for data pre-processing purposes to enhance the performance of the developed model through identifying the most relevant input variables (Poh et al. 2018). In other cases, feature selection can reduce the computational time of the forecasting model, such as when dealing with large dimensional datasets (Elkholosy 2020).

Two different algorithms are used to identify which attributes have a significant impact on the labor resource requirements. First, a deterministic algorithm is applied, which behaves predictably. The algorithm produces the same output every time, and the underlying machine passes through the same sequence of states. Then, a greedy algorithm is applied. This algorithm follows the problem-solving heuristic of making the locally optimal choice at each stage to find a global optimum. The proposed feature selection approach includes two deterministic greedy algorithms - 'forward selection' and 'backward elimination' - while some enhancements are added to the standard algorithms which are described in the following.

With an initial population with n attributes in the input dataset. Each attribute is dedicated into only one of the features. The attributes are evaluated by forward selection algorithm and only k number of best ones are selected. For each of the k attribute sets the algorithm would do: if there are j unused attributes, makes j copies of the attribute set and adds one of the previously unused attributes to the attribute set. This step repeats if the performance is improved.

Backward elimination is the opposite of the forward selection. The algorithm starts with an attribute set which uses all features. Then, evaluates all attribute sets and selects the best k . For each of the k attribute sets the algorithm would do: if there are j attributes used, makes j copies of

the attribute sets, and removes one of the previously unused attributes from the attribute set. The loop repeats as long as the performance is improved.

The method utilizes a machine learning algorithm to identify the best features trying different combinations. After applying the forward selection and backward elimination, 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 6 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 attributes increased to 27. Feature selection reduces the number of attributes by eliminating irrelevant attributes and improves the performance of the forecasting models.

Feature selection is applied to each of the clusters that were developed using the k-means algorithm. The selected features are evaluated using 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 and backward elimination are used in the wrapper method.

It is noteworthy that due to the limited attributes of the historical data, the feature selection approach could not be implemented. The wrapper method requires a bigger dataset with more input variables. Except the first cluster, other clusters did not have sufficient records for proper applying

of feature selection. Also, the first cluster had very limited input variables which made the feature selection inapplicable.

4.4 Forecasting Models

Different prediction algorithms are explored to forecast the labor hours required for a work package. Linear regression and Recurrent Neural Network (RNN) are the investigated algorithms for the defined problem. Linear model is used due to its simplicity and ability to predict with minimal error. RNN is one of the most powerful machine learning algorithms that are widely used in different industries. Also, it is very formidable in dealing with complex problems and nonlinear data.

The proposed model is trained by the collected historical data to identify inherent trends in the data and unseen relationships between the output and input features. The model was initially trained to forecast the required labor hours for a work package, using all work packages in the collected dataset. But, the model did not perform acceptable and the error value was high. This huge error was caused due to the wide range of work packages and their durations. Consequently, cluster analysis was carried out as the possible solution to group work packages that have similar durations. Finally, forecasting model was applied for each cluster, and its performance was compared to the initial trial that included all work packages.

By implementing feature selection process, the significant impacting factors were identified. These attributes are considered as the inputs for the forecasting models. They were used to train and test the forecasting models to predict the desired output. Baseline model, linear regression, and RNN models were investigated, and the performance was compared to determine which machine learning algorithm fits best for the defined problem.

4.4.1. Baseline Model

Before building a trainable model, it would be good to have a performance baseline as a point for comparison with the later more complicated models. This first task is to predict labor hour one time unit in the future given the current value of all features. The current values include the current labor hour. So, start with a model that just returns the current labor hour as the prediction, predicting "No change". In other words, the baseline model's prediction is simply giving the value of the previous time unit as the forecasted value for the next time unit. Obviously, the baseline model will work less well for making a prediction further in the future.

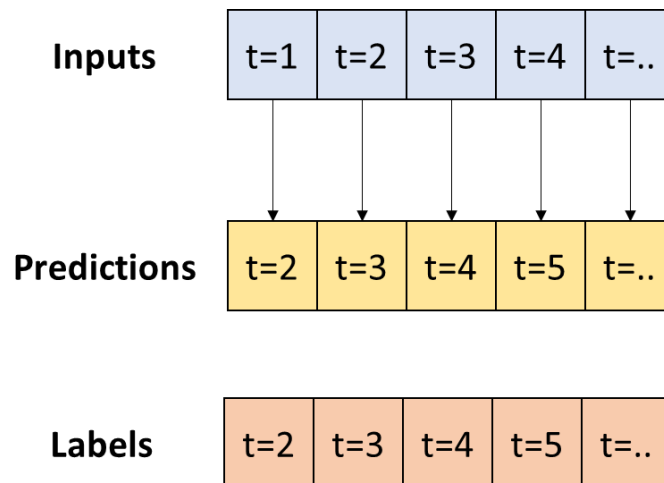


Figure 14. Baseline Model Schema

4.4.2. Linear Model

The simplest trainable model is to insert linear transformation between the input and output. A linear regression algorithm is developed for the model, and the performances is evaluated. The process of building linear regression models does not require many configurations in the 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 model requires numerical input variables, which are used to predict the labor hours by assigning weights

to every variable. The linear model does prediction of each time step independently from other time units which means that each time step has its own linear model.

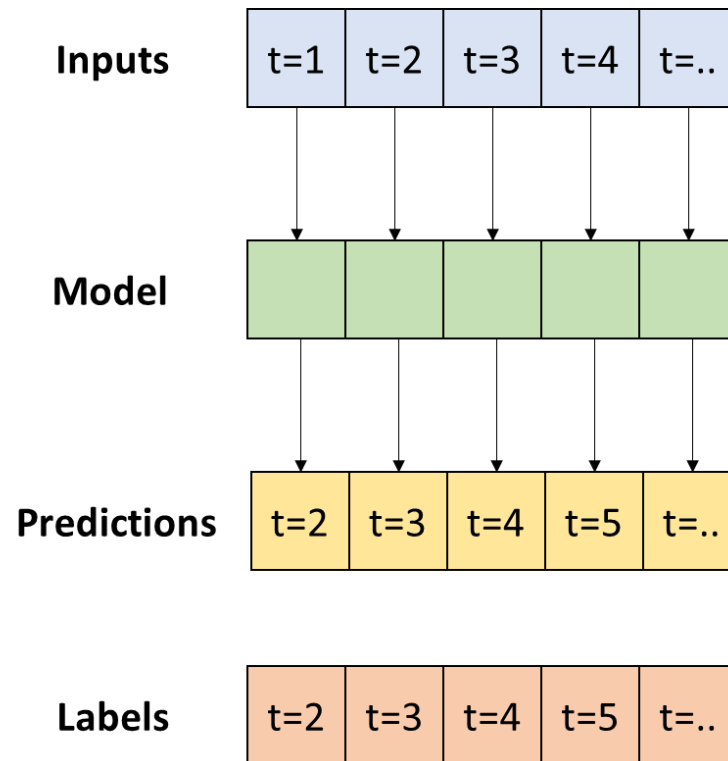


Figure 15. Linear Model Schema

4.4.3. Recurrent Neural Network Model

A single-time-step model like linear model has no context for the current values of its inputs. It cannot see how the input features are changing over time. To address this issue the model needs access to multiple time steps when making predictions (Figure 16).

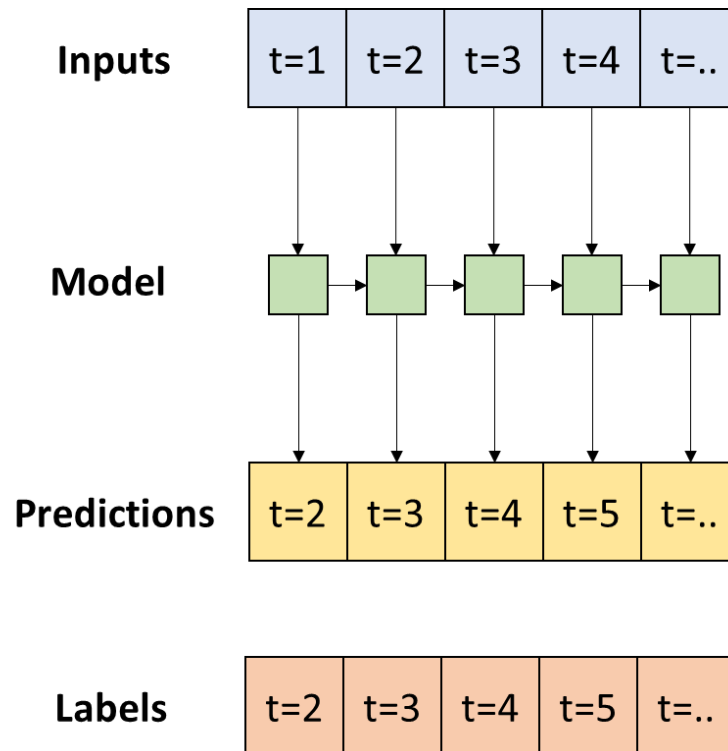


Figure 16. RNN Model Schema

Recurrent neural networks (RNN) are a class of neural networks that is powerful for modeling sequence data such as time series or natural language. What makes RNN so powerful is the fact that it does not take into consideration just the actual input but also the previous input which allows it to memorize what happens previously. They are distinguished by their “memory” as they take information from prior inputs to influence the current input and output. While traditional deep neural networks assume that inputs and outputs are independent of each other, the output of recurrent neural networks depend on the prior elements within the sequence. RNN unlike classic algorithms such as naïve Bayesian, works well on sequence data because it takes the time step i as input and combine with the output of time step $i-1$, the same thing would be applied for time step $i+1$ and this is the reason it is called recurrent neural network because clearly the neural network applies the same operations on each time step i of the sentence. 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.

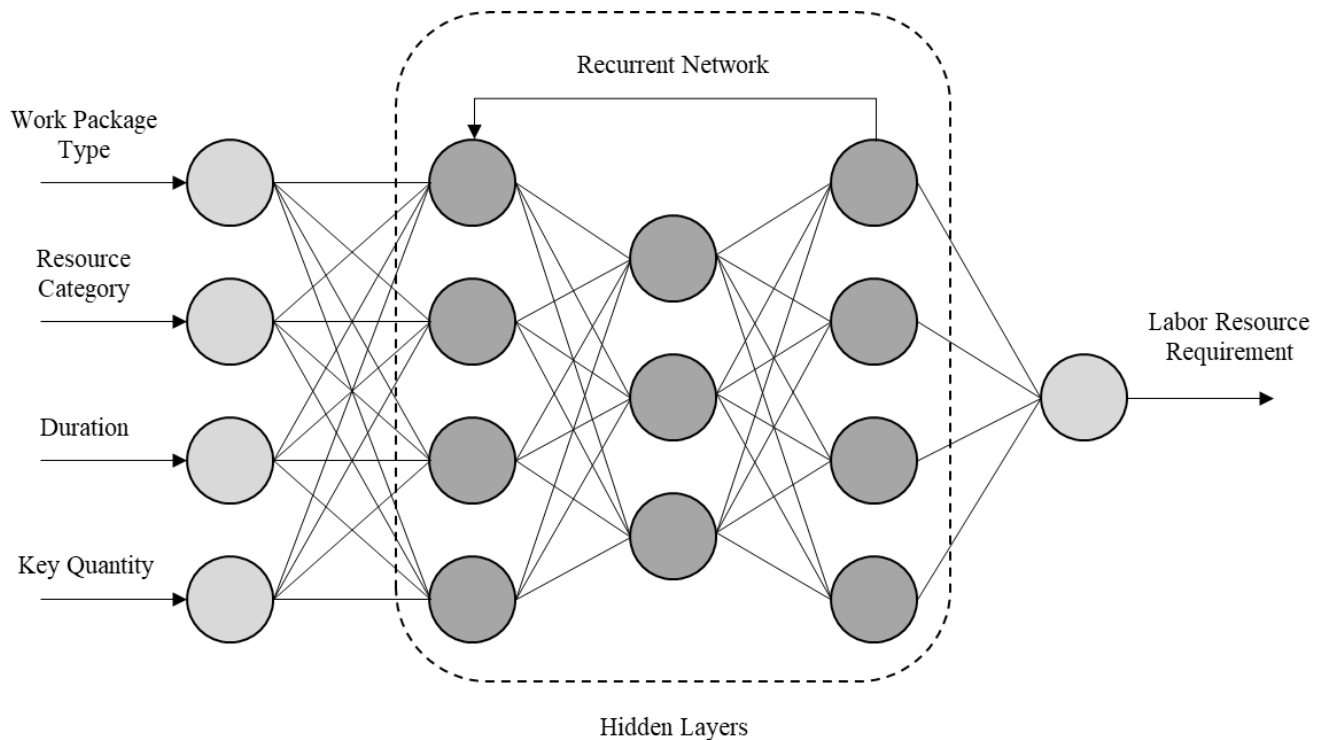


Figure 17. RNN Model Structure

Another distinguishing characteristic of recurrent networks is that they share parameters across each layer of the network. While feedforward networks have different weights across each node, recurrent neural networks share the same weight parameter within each layer of the network. That said, these weights are still adjusted in the through the processes of backpropagation and gradient descent to facilitate reinforcement learning.

As previously mentioned, the RNN does predictions step-by-step. This means that in our model, the algorithm is initially trained to forecast the required labor hours for the first day. Then the

network performs some enhancements automatically and is prepared for predicting the next day. This cycle takes place until the algorithm reaches the last day within the work package.

4.5 Performance Evaluation

Basically, the estimated performance of a model tells how well it performs on unseen data. Accordingly, after developing the forecasting model, model validation is required to ensure that the model works properly on unseen data. As mentioned before, there are two approaches that are used to evaluate the performance of the developed model. The first method involves splitting the dataset into training set, validation set and a testing set; with the split percentage of 70%, 20% and 10% for training, validating, and testing, respectively. The other technique is k-folds cross validation. This method splits the input data into k folds ($k=8$). The model is trained using the $k-1$ fold and tested with the remained fold. This process takes place until all the records in the dataset are included in the test set.

Besides, there are two common metrics used for evaluating the regression models namely, Root Mean Squared Error and Mean Absolute Error (Narula and Wellington 1977). The Mean Absolute Error (MAE) is the sum of the absolute differences between predictions and actual values (Equation 2). On the other hand, Root Mean Squared Error (RMSE) measures the average magnitude of the error by taking the square root of the average of squared differences between prediction and actual observation (Equation 3). The MAE is the main criterion used to measure the performance of the developed forecasting model.

$$MAE = \frac{1}{n} \times \sum_1^n |y_{pred} - y_{ref}| \text{ where } y_{pred} = \text{predicted value. } y_{ref} = \text{actual value} \quad (2)$$

$$RMSE = \sqrt{\sum_1^n \frac{(y_{pred} - y_{ref})^2}{n}} \text{ where } y_{pred} = \text{predicted value. } y_{ref} = \text{actual value} \quad (3)$$

Figure 18 indicates the MAE for the RNN model used in each cluster to forecast the labor hours. The forecasting model for the first cluster performed with lower error in comparison to the other clusters because the number of work packages used to train the models were greater than the number of projects in the other clusters.

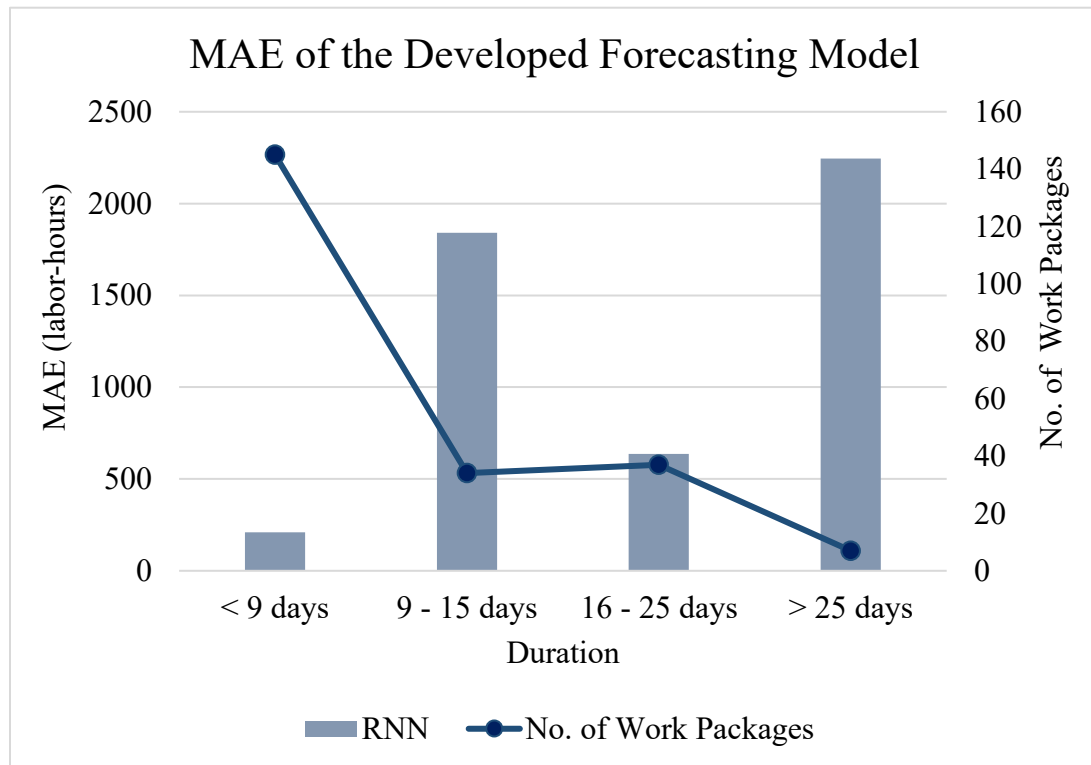


Figure 18. MAE of the Forecasting Model

The performance of the forecasting models was considerably enhanced after omitting outliers, applying cluster analysis, and selecting features prior to the development of the forecasting model. Proposing one single forecasting model for all work packages was not acceptable and in conflict with the initial problem definition of investigating the various durations and labor hours. With the high error value and deficient performance, the result was not reliable for an estimate from the

model. Thus, a forecasting model was used for each cluster resulting in a noticeable decrease in the error value.

After developing separate models, the errors of the models were still high, however they performed much better than the baseline model. This concludes that the proposed models can perform well in predicting the required labor hours as it is better than using the baseline model and estimates done by the industry practitioners. 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 Summary

In this chapter, the taken steps of developing the model from data collection to performance evaluation were discussed. As mentioned, data preparation was challenging as the data included too much coded information from different sources. The collected data had much missing values compared to the designed data acquisition model. Hence, the study was limited in different aspects. Then, outlier detection methods were applied to eliminate the outliers found in the data set. The cluster analysis also helped in detecting the anomalies.

Considering the wide variety of the work packages, developing one single model that can forecast any construction work package was not applicable. Therefore, the clustering technique (k-means) was applied to group the relevant work packages based on the duration. Then, a forecasting model was trained for every cluster.

Two different robust machine learning algorithms (linear regression and RNN) were evaluated to determine which algorithm is the best possible choice for the purpose of this study. Finally, the RNN algorithm was chosen due to its high capability in forecasting time series and dealing with anomalies. The MAE of the model was nearly high due to the limited data for training, but the performance would be still acceptable and can be easily improved by adding more training data.

The error value for the first cluster which includes the work packages less than 9 days, has the least error among the other clusters. The lower error is due to the higher number of training records. Although, the overall mean absolute error of the model is nearly 4,900 labor-hours, but it is still acceptable in comparison to the current industry practices. This is caused by the limited number and diversity of work packages used for training and the lack of available information as input variables. The accuracy of the models is expected to increase when the data is collected based on the proposed data acquisition model, and by increasing the number of work packages to train the RNN model.

Chapter 5 Implementation of the Labor Resource Forecasting Model

5.1 Introduction

The construction companies mostly do not take the advantages of their historical data. Besides, professional complex software can be complicated and confusing. Having a simple forecasting application can easily tackle these issues.

During this research, a simple application for estimating the labor hours for construction work packages is developed. The application utilizes the developed forecasting model and provides the prediction based on the user's inputs. Figure 19 shows the interface of the application. The interface is designed in a way to provide a simple and straight forward service into the user whether it is a project manager or a researcher. Also, the output is presented decently so that it can be easily interpreted and utilized. Developing such computer application can bring many advantages including: (1) easy to run and no need for complex installation procedure. (2) providing a user-friendly interface for utilizing the developed model. (3) Presenting a framework to aid project analysis and realistic future planning. (4) Avoiding users interfere with the back-end codes. (5) Providing a decent visualization for the output of the model.

Curve Forecasting Model

Enter the information :

Work Package Category	Resource Type	Duration
<input type="button" value="Choose"/>	<input type="button" value="Choose"/>	<input type="button" value="Choose"/>

Type the Key Quantity:

Figure 19. Interface of the Application

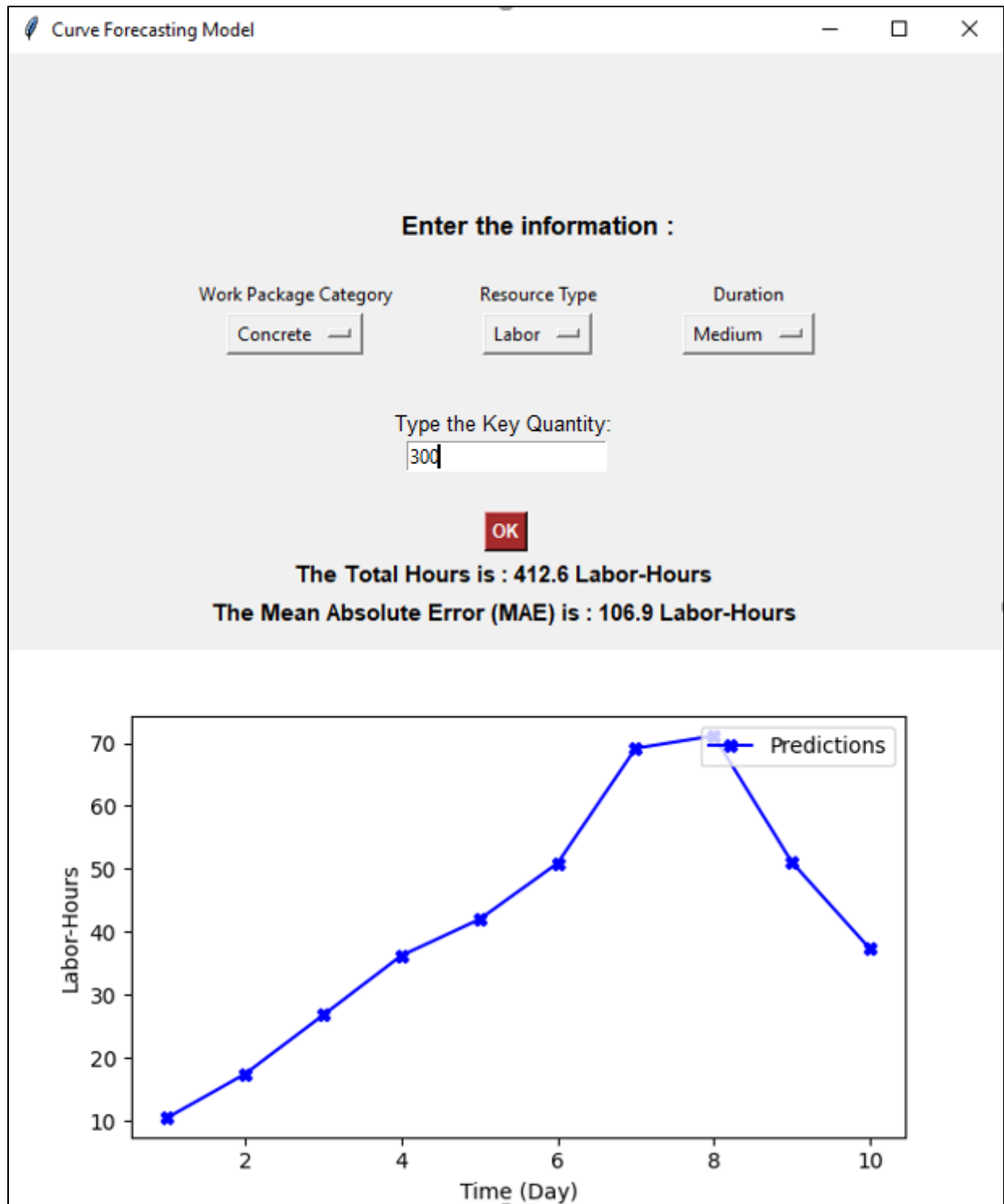


Figure 20. Sample Run of the Application

5.2 Application Components

The development of the application was an updating process. Figure 21 presents the initial version of the interface. The feedback from experienced project managers provided useful modifications to make the application useable for industry experts. The initial idea was to develop a generic application to support different work packages and resources. But limitations in time and data prevented developing such generic application and the current version is limited to the collected data from one contractor.

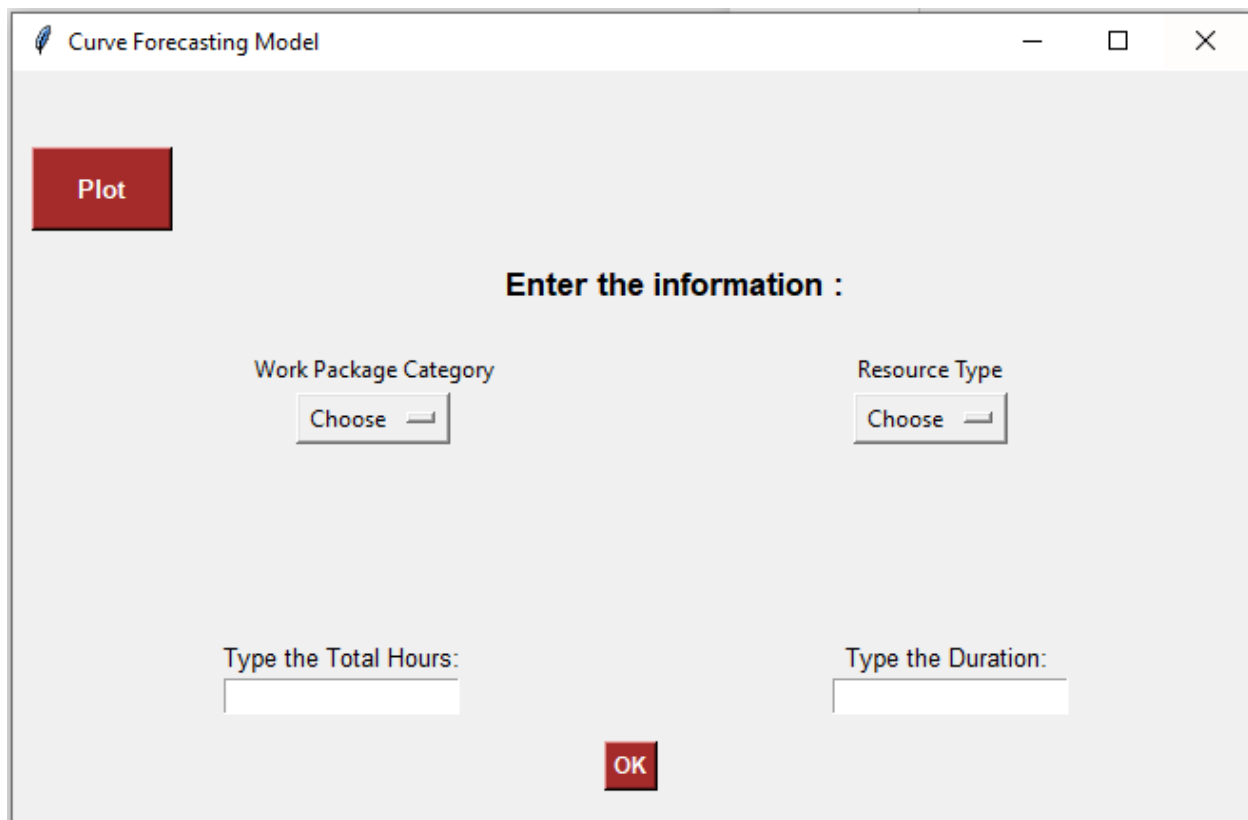


Figure 21. Initial Version of the Interface

The application includes two main parts. Following elaborates these two major components of the current application:

1. The main window and dropdown buttons (Figure 22):

To predict the labor resource requirements, the user selects the required inputs. After getting user's inputs, the application runs the trained RNN model.

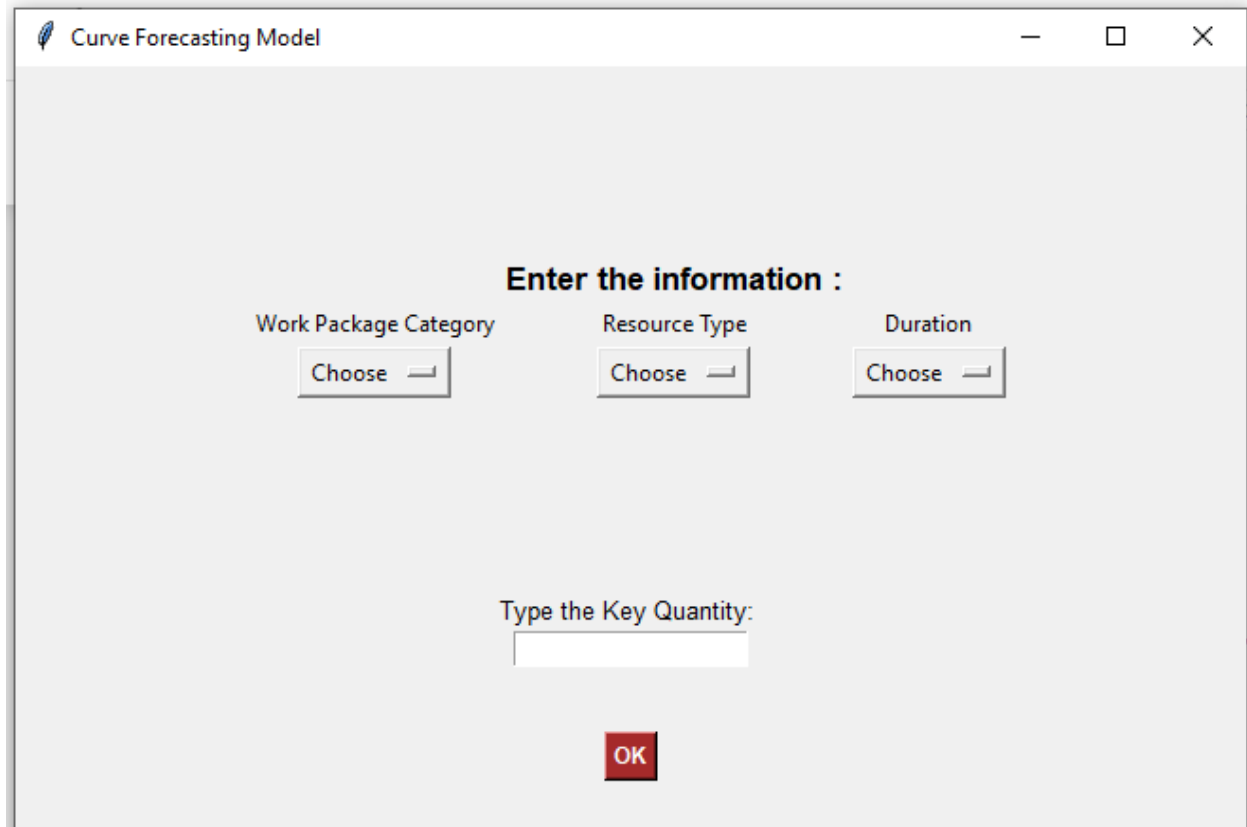


Figure 22. Main Window and Dropdown Buttons (Inputs)

There are 3 dropdown buttons for the work package category, resource, and duration. Figures 23, 24 and 25 represent these dropdowns and their setup codes.

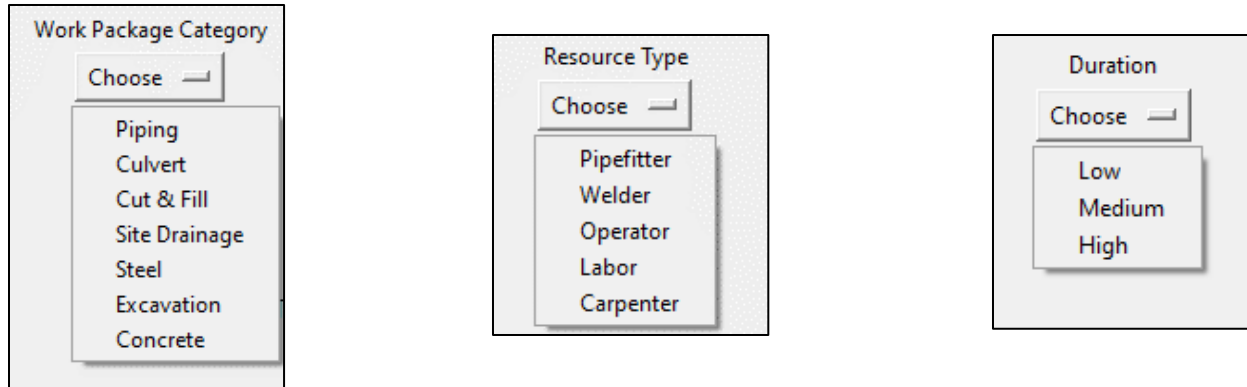


Figure 23. Dropdown Buttons and Lists

```

root = Tk()
root.title("Curve Forecasting Model")
# Add a grid
mainframe = Frame(root)
mainframe.grid(column=0, row=0, sticky=(N,W,E,S) )
mainframe.columnconfigure(0, weight = 1)
mainframe.rowconfigure(0, weight = 1)
mainframe.pack(pady = 100, padx = 100)
root.geometry("650x750")

```

Figure 24. Main Frame Setup Code

It is noteworthy that the categories of work packages and resources are limited to the construction ones and the other types such as management, planning and procurement are deleted from the lists. This is due to the different characteristic of these work packages. They usually last much longer in comparison to the construction work packages and has consistent labor hours spent each day. For example, a project management work package would last 200 days with a steady labor hour for each day while an excavation work package lasts mostly less than 30 days with a fluctuating required labor hours during the execution. As a result, to achieve a more accurate and beneficial model, the general work packages and resources are eliminated from the model and only construction ones are included in the proposed application.

```

# Create a Tkinter variable
tkvar1 = StringVar(root)
tkvar2 = StringVar(root)
tkvar3 = StringVar(root)
#tkvar4 = StringVar(root)

# Dictionary with options
choices1 = { 'Bolt Up', 'Concrete', 'Control Valve', 'Culvert', 'Cut & Fill',
             'Excavation', 'General', 'Piping', 'Project Management',
             'Rework', 'Services', 'Site Drainage', 'Steel', 'Superintend' }
choices2 = { 'Project Management', 'Overburden Operator', 'Carpenter', 'Welder',
             'HSE', 'Ironworker', 'Labor', 'Operator', 'Pipefitter', 'Project Exec',
             'Quality', 'Accounting' }
choices3 = { 'High', 'Medium', 'Low' }
#choices4 = { 'Construction Management', 'Design Bid Build', 'Design Build', 'In-Ho

tkvar1.set('Choose') # set the default option
tkvar2.set('Choose')
tkvar3.set('Choose')
#tkvar4.set('Choose')

Label(mainframe, text="Enter the information :", font=('helvetica', 13, 'bold'))

popupMenu = OptionMenu(mainframe, tkvar1, *choices1)
Label(mainframe, text="Work Package Category").grid(row = 3, column = 1)
popupMenu.grid(row = 4, column = 1)

popupMenu = OptionMenu(mainframe, tkvar2, *choices2)
Label(mainframe, text="Resource Type").grid(row = 3, column = 3)
popupMenu.grid(row = 4, column = 3)

popupMenu = OptionMenu(mainframe, tkvar3, *choices3)
Label(mainframe, text="Duration").grid(row = 3, column = 5)
popupMenu.grid(row = 4, column = 5)

Label(mainframe, text=" ").grid(row = 1, column = 3)

button = Button(root, text="OK", command = change_dropdown, bg='brown', fg='white')
button.pack()

```

Figure 25. Dropdown Buttons Setup Code

2. The output and plotting:

As previously discussed, a forecasting model is developed for different work packages clustered based on their durations. One of the trained recurrent neural networks is used to forecast the labor resource requirement of a new work package, and the selection of the cluster

is made when the user selects the duration. To avoid confusion, the codes are not accessible within application and only developers have access to modify the models and their parameters.

Figures 26 and 27 represent the sample codes for the output setup.

```
canvas2 = ttk.Canvas(root, width = 600, height = 60, relief = 'raised')
canvas2.pack()
label4 = ttk.Label(root, text="The Mean Absolute Error (MAE) is : " + str(ro
                    font=('helvetica', 11, 'bold'))
canvas2.create_window(300, 40, window=label4)

label5 = ttk.Label(root, text="The Total Hours is : " + str(round(TotalHours
                    font=('helvetica', 11, 'bold'))
canvas2.create_window(300, 15, window=label5)
```

Figure 26. Model's Output Setup Code

```
fig = Figure(figsize = (12, 8), dpi = 100)
x=np.arange(1,int(duration)+1,1)
y=PredictedHours
plot1 = fig.add_subplot(111)
plot1.plot(x,y , marker='X', label='Predictions', color='blue')
plot1.legend(loc="upper right")
plot1.set_xlabel('Time (Day)')
plot1.set_ylabel('Labor-Hours')
plot1.LineWidth = 4
canvas = FigureCanvasTkAgg(fig, master = root)
canvas.draw()
canvas.get_tk_widget().pack()

root.mainloop()
```

Figure 27. Output Plotting Setup Code

The model provides the user with various outputs including the graph of labor hours predicted for each working day by the RNN model, the total labor hours estimated for the work package, and the error value of the predicted hours which can be interpreted as a confidence interval. The output would provide a precise perspective about the labor hours needed for a given work package. For instance, Figure 28 represents the required duration and labor hours for a short

steel work package. According to the result, the work package needs 4 days for completion and a total of 162.5 labor hours with the MAE value equal to 6 labor hours.

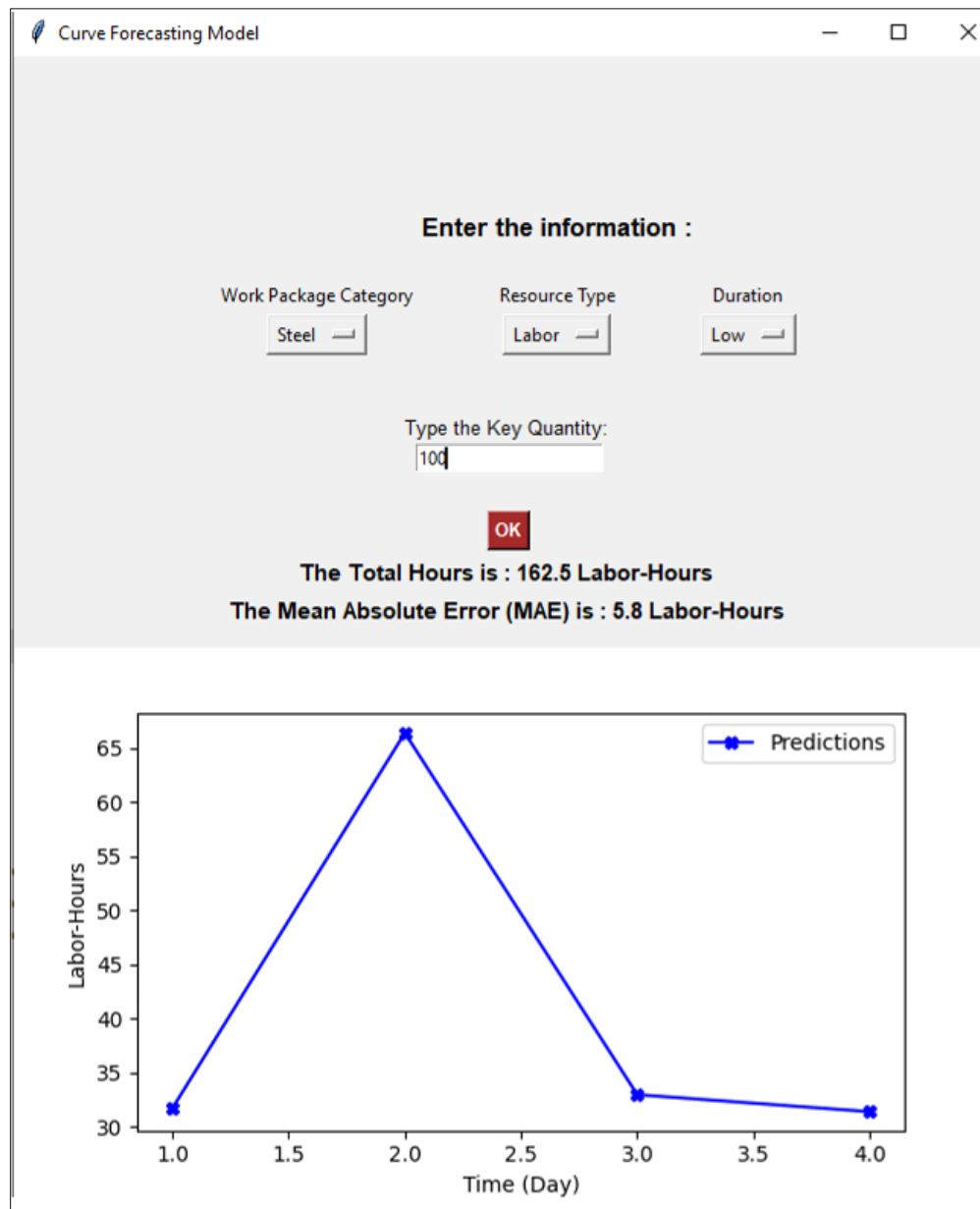


Figure 28. Model's Prediction for Required Labor Resource in a Steel Work Package

In another example, Figure 29 shows the predicted labor hours for a piping work package with the duration of 17 days while Figure 30 elaborates the real labor hours spent for a piping work package with similar characteristics.

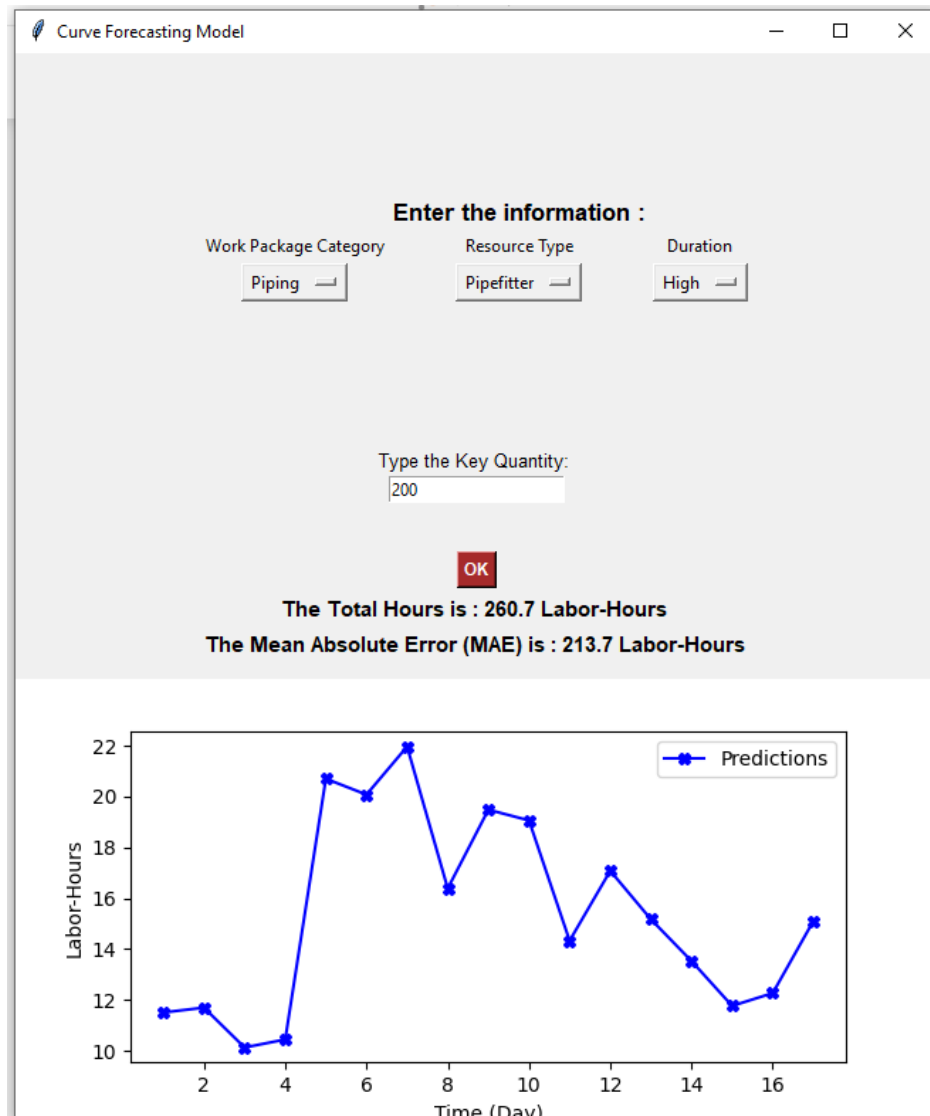


Figure 29. Predicted Labor Hours Required for a Piping Work Package

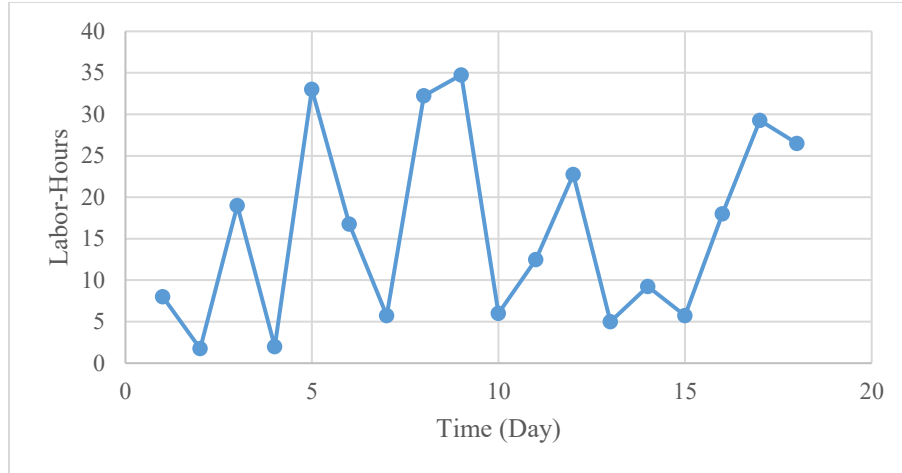


Figure 30. Actual Labor-Hours Spent for a Piping Work Package

As shown in Figure 29, the required hours for a piping work package with the duration of 17 days is predicted as 260 labor-hours. On the other hand, Figure 30 shows the spent labor hours for a similar piping work package which lasted 18 days. The actual total hours is 288 labor-hours which is near to the total o predicted hours. This comparison indicates that the developed application works properly, and its accuracy is acceptable. But it is clear that the two curves are not similar and in some of the time steps the difference between the actual and predicted hours is high. The reason of this error can be found in the inconsistency of the training dataset and the high variance in the actual labor hours spent in each time step which can lead to the confusion of the model and accordingly inaccurate estimation of the required labor hours for each day. By implementing the training process of the application with the proposed data acquisition model, such inconsistencies are reduced, and the performance is expected to improve.

Chapter 6 Summary, Limitations and Future Work

6.1 Research Summary

Considering the high rates of labor resources in construction projects, proposing a decent labor resource management framework is critical. Predicting needed labor requirements can be a substantial step towards labor resource management. With the recent developments in the area of artificial intelligence and machine learning, these technologies can potentially be adopted to develop resource prediction models. As the labor hours are often documented and recorded in historical datasets, there is a potential to use these datasets to analyze labor resources, identify underlying patterns, and forecast future trends.

Accordingly, this research aims to combine the benefits of artificial intelligence and historical data of previous projects to identify the significant factors affecting the labor hour requirements and to develop a generic forecasting model. This objective is achieved by using machine learning algorithms and developing neural network models for various construction work packages. Although, the error of the models is high due to the wide range of work packages and the limited number of records used to train the models.

In the first step, the key factors impacting the labor resource requirements were explored through previous studies and industry practices. Therefore, an initial list of attributes was provided. Then, historical data of previously performed construction projects was collected focusing on the explored key attributes. The proposed list of features consisted of the following attributes: work package type, resource category, duration, cost, complexity, and key quantity.

The process of data collection was challenging since some of the target attributes were not available by the contractors. The data was extracted from the tracking spreadsheets used by project

managers. These pre-defined spreadsheets included the information of the tasks and activities happening each day. But all the information was not accessible due to confidentiality. As a result, the collected dataset was limited into a small portion of the original dataset.

In the next step, the collected data was pre-processed and cleaned to be used as the training set for the models. Two machine learning algorithms (linear regression and RNN) were evaluated for using in the forecasting models. It was concluded that the recurrent neural network algorithm is more capable in predicting time series and was chosen as the final algorithm. Because of the wide range of the work packages, developing one forecasting model for all of them was not applicable and would lead to high error in estimated labor hours (Table 1). Hence, clustering analysis was performed on the dataset to group work packages based on their durations (Figure 31). Then, forecasting models were developed for each cluster. The comparison of the results (Table 1 and 3) clearly indicated that clustering has reduced the error value and increased the accuracy of the model. Also, a feature selection method was designed to identify the most impacting factors. But due to the limited attributes of the collected dataset, the feature selection was not applied.

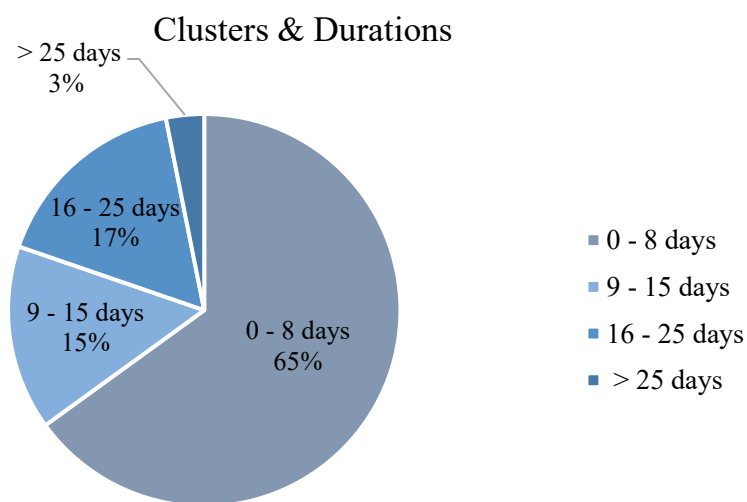


Figure 31. Cluster Analysis Result

The performance of the RNN model was evaluated and the results is presented in Table 3. The developed model for the first cluster – work packages with less than 9 days -had the lowest error among the other clusters due to the efficient number of the training records. The mean absolute error of the cluster is 209 labor-hours which is still high but can be improved by adding high-quality training dataset. The high value for the error is mainly due to the limited number of available records and the lack of some attributes as the models' inputs. The accuracy would be improved by increasing the training dataset records and collecting information according to the provided data acquisition model.

Table 3. RNN Model Clusters and Results

Cluster	Number of Work Packages	Duration (day)	MAE (labor-hours)
1	145	0 - 8	± 209
2	34	9 - 15	± 1841
3	37	16 - 25	± 636
4	7	≥ 26	± 2245

The output of the developed models is the daily labor resource requirements estimated for a specific type of work package. The result of the model is not a single value but the distribution of the required labor hours over the work package duration. Consequently, the output would be more reliable and project managers could benefit more from the result as it is not limited into only an estimated total hour for the whole work package.

Eventually, a computer application was designed for the proposed model. The application was developed to provide a user-friendly interface, so that the user – whether a senior project manager

or a researcher - can easily benefit from the developed forecasting model. By running the application, the user can select the desired inputs and after running the trained model at the back end, the predicted result is plotted with the estimated error value.

6.2 Limitations

While this research aims to develop a forecasting model for labor resource requirements, there were some limitations which can be summarized as follows:

The most important during this research was the data collection process. Finding the appropriate contractor was challenging as there are not many contractors who are willing to share their historical data due to confidential issues and their policies. Also, the pandemic and lockdown situation highly impacted the communications and reaching out to industry partners became difficult. This situation caused delay in the data collection process.

After finding the contractor only a limited dataset was received. The shared dataset was including only few attributes of only one construction project. This small dataset did not allow this research to fully benefit from the proposed approach. The proposed feature selection method could not be applied since the input variables were limited. Moreover, the input variables were limited to 4 attributes as the other expected inputs were not available. All these issues lead to decreased accuracy in the performance of the prediction models.

The quality of the training dataset was not of high quality. As mentioned, the dataset was limited to the construction projects with similar characteristics including only specific types of work packages. Also, some of their work packages were general ones which had long durations with constant daily labor hours. Such work packages would disturb the training process and cause inaccurate model.

The machine learning algorithms especially RNN, need very high number of records for training in order to perform predictions with low error. Otherwise, they are not capable to forecast the labor hours precisely and cannot be reliable. Consequently, the performance of the developed model is impacted and needs much more high-quality historical data for training to improve its high error predictions.

6.3 Future Work and Recommendations

This research can be expanded in different directions. The current model can be expanded to forecast more work package types and to consider other resource categories. Collecting more historical data from different construction projects and supplying high-quality dataset to the model can improve the performance and reliability. Also, adding more significant factors as inputs to the model can potentially increase the accuracy.

The developed model can be integrated into a greater model which is capable to forecast different aspects of a work package such as duration, required labor resources, key quantity etc. With the rapid advancement of artificial intelligence, new algorithms with more capabilities in forecasting are emerging. RNN is just a sample of these brand-new algorithms which was used in this research. The new algorithms and techniques are always potential options for improvement of the current models.

References

- Agapiou, A., Price, A. D., & McCaffer, R. (1995). Planning future construction skill requirements: understanding labour resource issues. *Construction Management and Economics*, 13(2), 149-161.
- Alizadeh, S. S., Mortazavi, S. B., & Mehdi Sepehri, M. (2015). Assessment of accident severity in the construction industry using the Bayesian theorem. *International journal of occupational safety and ergonomics*, 21(4), 551-557.
- 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.
- Amrutha, V. N., & Geetha, S. N. (2020). A systematic review on green human resource management: Implications for social sustainability. *Journal of Cleaner Production*, 247, 119131.
- and Computer Engineering. IEEE. p. 1–6
- Antonioli, D., Mancinelli, S., Mazzanti, M. (2013). Is environmental innovation embedded within high-performance organisational changes? The role of human resource management and complementarity in green business strategies. *Res. Policy* 42 (4), 975–988.
- Apaydin, H., Feizi, H., Sattari, M. T., Colak, M. S., Shamshirband, S., & Chau, K. W. (2020). Comparative analysis of recurrent neural network architectures for reservoir inflow forecasting. *Water*, 12(5), 1500.
- Arnett, D. B., Laverie, D. A., & McLane, C. (2002). Using job satisfaction and pride as internal-marketing tools. *Cornell hotel and restaurant administration quarterly*, 43(2), 87-96.
- Baiden, B.K., Price, A.D., 2011. The effect of integration on project delivery team effectiveness. *Int. J. Proj. Manag.* 29 (2), 129–136.
- Ballesteros Pérez, P., González-Cruz, M^a.C., Pastor-Ferrando, J.P. (2010). Analysis of construction projects by means of value curves. *Int. J. Proj. Manag.* Volume 28 (Issue 7), 719–731.

- Baloh, P., Desouza, K.C., Hackney, R. (2012). Contextualizing organizational interventions of knowledge management systems: a design science perspective. *J. Am. Soc. Inf. Sci. Technol.* 63 (5), 948–966.
- Partington, D., Pellegrinelli, S., Young, M., 2005. Attributes and levels of programme management competence: an interpretive study. *Int. J. Proj. Manag.* 23 (2), 87–95.
- Baron, R. A. (2003). Human resource management and entrepreneurship: Some reciprocal benefits of closer links. *Human Resource Management Review*, 2(13), 253-256.
- Becker, B., & Gerhart, B. (1996). The impact of human resource management on organizational performance: Progress and prospects. *Academy of management journal*, 39(4), 779-801.
- Becker, B.E., Huselid, M.A. (2006). Strategic human resources management: where do we go from here? *J. Manag.* 32 (6), 898–925.
- Belout, A., Gauvreau, C. (2004). Factors influencing project success: the impact of human resource management. *Int. J. Proj. Manag.* 22 (1), 1–11.
- Bonaccorso, G. (2018). *Machine Learning Algorithms: Popular algorithms for data science and machine learning*. Packt Publishing Ltd.
- Bourne, L., Walker, D.H. (2005). The paradox of project control. *Team Perform. Manag.* 11 (5–6), 157–178.
- Bowen, D.E., Ostroff, C. (2004). Understanding HRM-firm performance linkages: the role of the "Strength" of the HRM system. *Acad. Manag. Rev.* 29 (2), 203–221.
- Bredin, K. (2008). People capability of project-based organizations: a conceptual framework. *Int. J. Proj. Manag.* 26 (5), 566–576.
- Briscoe, G., & Wilson, R. A. (1993). *Employment forecasting in the construction industry*. Ashgate Publishing.
- Buller, P.F., McEvoy, G.M., 2012. Strategy, human resource management and performance: sharpening line of sight. *Hum. Resour. Manag. Rev.* 22 (1), 43–56.
- Bush, V. G. (1973). *Construction management: A handbook for contractors, architects, and students*. Reston Publishing.

- Campion, M.A., Medsker, G.J., Higgs, A.C. (1993). Relations between work group characteristics and effectiveness: implications for designing effective work groups. *Pers. Psychol.* 46 (4), 823–847
- Chan, A.P.C., Chiang, Y.H., Mak, S.W.K., Choy, L.H.T. and Wong, J.M.W. (2006) Forecasting the demand for construction skills in Hong Kong, *Construction Innovation*, 6(1), 3-19.
- Chan, A.P.C., Wong, J.M.W. and Chiang, Y.H. (2003) Modelling labour demand at project level – an empirical study in Hong Kong, *Journal of Engineering, Design and Technology*, 1(2), 135-50.
- Chen, L. H., Liaw, S. Y., & Lee, T. Z. (2003). Using an HRM pattern approach to examine the productivity of manufacturing firms—an empirical study. *International Journal of Manpower*.
- Chiang, Y. H., Wong, F. K. W., & Liang, S. (2018). Fatal construction accidents in Hong Kong. *Journal of construction engineering and management*, 144(3), 04017121.
- Chukwu, A. U., & Adepoju, K. A. (2012). On the power efficiency of artificial neural network (ANN) and the classical regression model. *Progress in Applied Mathematics*, 3(2), 28-34.
- Cleland, D.I., Ireland, L.R. (2006). *Project Management: Strategic Design and Implementation*, 5th ed. McGraw-Hill Professional, New York.
- Dang, Y., Zhang, Y., Zhang, D., & Zhao, L. (2005, August). A KNN-based learning method for biology species categorization. In *International Conference on Natural Computation* (pp. 956-964). Springer, Berlin, Heidelberg.
- Datta, D.K., Guthrie, J.P., Wright, P.M. (2005). Human resource management and labor productivity: does industry matter? *Acad. Manag. J.* 48 (1), 135–145.
- Davis, K. (2014). Different stakeholder groups and their perceptions of project success. *Int. J. Proj. Manag* 32 (2), 189–201.
- Dessler, G. (2012). *A Framework for Human Resource Management*, 7th ed. Prentice Hall, Pearson.

- Druker, J., White, G., Hegewisch, A., & Mayne, L. (1996). Between hard and soft HRM: human resource management in the construction industry. *Construction Management & Economics*, 14(5), 405-416.
- Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification*. John Wiley & Sons. Inc., New York, 2.
- Ehrenberg, R. G., & Smith, R. S. (2003). *Modern Labor Economics: theory and public policy* eight edition.
- Franco-Lopez, H., Ek, A. R., & Bauer, M. E. (2001). Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. *Remote sensing of Environment*, 77(3), 251-274.
- Ganesan, S., Hall, G., and Chiang, Y. H. (1996). *Construction in Hong Kong: Issues in labor supply and technology transfer*, Aldershot, Avebury, U.K.
- Garavan, T. N., & Morley, M. J. (2006). Re-dimensionalising boundaries in the theory and practice of Human Resource Development. *International Journal of Learning and Intellectual Capital*, 3(1), 3-13.
- Gerassis, S., Martín, J. E., García, J. T., Saavedra, A., & Taboada, J. (2017). Bayesian decision tool for the analysis of occupational accidents in the construction of embankments. *Journal of construction engineering and management*, 143(2), 04016093.
- Gershenson, C. (2003). Artificial neural networks for beginners. arXiv preprint cs/0308031.
- Gidado, K. I., & Millar, A. J. (1992, September). The effect of simple overlap of the stages of building construction on the project complexity and contract time. In *Proceedings of the 8 th Annual Conference, Association of Researchers in Construction Management* (pp. 307-317).
- Gondia, A., Siam, A., El-Dakhakhni, W., & Nassar, A. H. (2020). Machine Learning Algorithms for Construction Projects Delay Risk Prediction. *Journal of Construction Engineering and Management*, 146(1), 04019085.
- Gorunescu, F. (2011). *Data Mining*, San Francisco, CA, itd. Morgan Kaufmann. doi, 10, 978-3.
- Gould, F. E. (2002). *Construction project management*, Prentice-Hall, Upper Saddle River, N.J.

- Gruneberg, S.L. (1997), *Construction Economics – An Introduction*, Macmillan, London
- Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003, November). KNN model-based approach in classification. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (pp. 986-996). Springer, Berlin, Heidelberg.
- Gurmu, A. T., & Ongkowijoyo, C. S. (2020). Predicting Construction Labor Productivity Based on Implementation Levels of Human Resource Management Practices. *Journal of Construction Engineering and Management*, 146(3), 04019115.
- Hammad, A. M. (2009). An integrated framework for managing labour resources data in industrial construction projects: A Knowledge Discovery in Data (KDD) approach (Vol. 71, No. 01).
- Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.
- Handy, C. B. (1985). *Understanding organizations*, 3rd Ed., Penguin, Harmondsworth, U.K
- Fairweather, V. (1986). "Record high-rise, record low steel." *Civ. Eng. (N.Y.)*, 56(8), 42–45
- Haykin, S. (1999). *Learning processes; single-layer perceptrons; multilayer perceptrons. Neural Networks A Comprehensive Foundation*. 2nd ed. USA: Prentice Hall International Inc, 14-68.
- Haykin, S. (2007). *Neural networks: a comprehensive foundation*. Prentice-Hall, Inc.
- Hendry, D.F. (1985), *Empirical Modelling in Dynamic Econometrics*, Nuffield College, Oxford.
- Heravi, G., & Eslamdoost, E. (2015). Applying artificial neural networks for measuring and predicting construction-labor productivity. *Journal of Construction Engineering and Management*, 141(10), 04015032.
- Hevner, A.R. (2007). The three-cycle view of design science research. *Scand. J. Inf. Syst.* 19 (2), 87.
- Hibat-Allah, M., Ganahl, M., Hayward, L. E., Melko, R. G., & Carrasquilla, J. (2020). Recurrent neural network wave functions. *Physical Review Research*, 2(2), 023358.
- Huang, C. H., & Hsieh, S. H. (2020). Predicting BIM labor cost with random forest and simple linear regression. *Automation in Construction*, 118, 103280.

- Huemann, M. (2010). Considering human resource management when developing a project-oriented company: case study of a telecommunication company. *Int. J. Proj. Manag.* 28 (4), 361–369.
- Huselid, M. A. (1995). The impact of human resource management practices on turnover, productivity, and corporate financial performance. *Academy of management journal*, 38(3), 635-672.
- Inyang, B. J. (2011). Creating value through people: Best human resource (HR) practices in Nigeria. *International business and management*, 2(1), 141-150.
- Iwu, C. G. (2016). Effects of the use of electronic human resource management (E-HRM) within human resource management (HRM) functions at universities. *Acta Universitatis Danubius. Administratio*, 8(1).
- Jantan, H., Hamdan, A. R., & Othman, Z. A. (2009). Knowledge discovery techniques for talent forecasting in human resource application. *World Academy of Science, Engineering and Technology*, 50, 775-783.
- Katz, H. C., Kochan, T. A., Keefe, J. H., Lazear, E., & Eads, G. C. (1987). Industrial relations and productivity in the US automobile industry. *Brookings papers on economic activity*, 1987(3), 685-727.
- Kubat, M., & Cooperson Jr, M. (2001). A reduction technique for nearest-neighbor classification: Small groups of examples. *Intelligent Data Analysis*, 5(6), 463-476.
- Kukenberger, M.R., Mathieu, J.E., Ruddy, T. (2012). A cross-level test of empowerment and process influences on members' informal learning and team commitment. *J. Manag.*
- Lall, U., & Sharma, A. (1996). A nearest neighbor bootstrap for resampling hydrologic time series. *Water resources research*, 32(3), 679-693.
- Laufer, A., Woodward, H., & Howell, G. A. (1999). Managing the decision-making process during project planning. *Journal of Management in Engineering*, 15(2), 79-84.

- Lee, B. H., & Scholz, M. (2006). A comparative study: Prediction of constructed treatment wetland performance with k-nearest neighbors and neural networks. *Water, Air, and Soil Pollution*, 174(1-4), 279-301.
- Lemessany, J., & CLAPP, M. A. (1978). Resource inputs to construction: the labour requirements of house building. Building Research Establishment.
- Li, H. (1996). Case-based reasoning for intelligent support of construction negotiation. *Information & Management*, 30(5), 231-238.
- Lim, C.S., Mohamed, M.Z. (1999). Criteria of project success: an exploratory reexamination. *Int. J. Proj. Manag.* 17 (4), 243–248.
- Liu, M., & Ballard, G. (2008, July). Improving labor productivity through production control. In *Proceedings of the 11th Annual Conference of International Group for Lean Construction*.
- Loosemore, M., Dainty, A., & Lingard, H. (2003). Human resource management in construction projects: strategic and operational approaches. Taylor & Francis.
- Marchington, M., & Wilkinson, A. (2002). People management and development: Human resource management at work. CIPD Publishing.
- McClelland, J. L., Rumelhart, D. E., & PDP Research Group. (1986). Parallel distributed processing. *Explorations in the Microstructure of Cognition*, 2, 216-271.
- McConnell, C. R., Brue, S. L., and Macpherson, D. A. (2003). Contemporary labor economics, McGraw-Hill, London.
- Meharie, M. G., & Shaik, N. (2020). Predicting Highway Construction Costs: Comparison of the Performance of Random Forest, Neural Network and Support Vector Machine Models. *Journal of Soft Computing in Civil Engineering*, 4(2), 103-112.
- Mourya, S. K., & Gupta, S. (2012). Data mining and data warehousing. Alpha Science International, Ltd.
- Muizz O. Sanni-Anibire, Rosli Mohamad Zin & Sunday Olusanya Olatunji (2020): Machine learning model for delay risk assessment in tall building projects, *International Journal of Construction Management*

- Mutua, A. M. (2019). The Impact of Human Resource Planning On Organizational Performance (Doctoral dissertation, University of Nairobi).
- Nakanishi, Y. (2001). Dynamic labor demand using error correction model. *Applied Economics*, 33(6), 783-790.
- Paliwal, M., & Kumar, U. A. (2009). Neural networks and statistical techniques: A review of applications. *Expert systems with applications*, 36(1), 2-17.
- Pant, I., Baroudi, B. (2008). Project management education: the human skills imperative. *Int. J. Proj. Manag.* 26 (2), 124–128.
- Patterson, D. (1996) *Artificial Neural Networks*. Prentice Hall, Singapore.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12, 2825-2830.
- 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.
- Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, 4(2), 1883.
- Pfeffer, J. (1998). *The Human Equation: Building Profits by Putting People First*. Harvard Business Review Press, Boston.
- Pinker, E. J., & Larson, R. C. (2003). Optimizing the use of contingent labor when demand is uncertain. *European Journal of Operational Research*, 144(1), 39-55.
- Pinto, J.K., Prescott, J.E. (1988). Variations in critical success factors over the stages in the project life cycle. *J. Manag.* 14 (1), 5–18.
- Poh, C. Q., Ubeynarayana, C. U., & Goh, Y. M. (2018). Safety leading indicators for construction sites: A machine learning approach. *Automation in construction*, 93, 375-386.
- Pournader, M., Tabassi, A. A., & Baloh, P. (2015). A three-step design science approach to develop a novel human resource-planning framework in projects: the cases of construction projects in USA, Europe, and Iran. *International journal of project management*, 33(2), 419-434.

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3), e1301.

Project Management Institute, (2013). *A Guide to the Project Management Body of Knowledge (PMBOK Guide)*, 5th ed. Project Management Institute, Newtown Square, Pennsylvania

Raiden, A.B., Dainty, A.R., Neale, R.H., (2004). Current barriers and possible solutions to effective project team formation and deployment within a large construction organisation. *Int. J. Proj. Manag.* 22 (4), 309–316.

Reilly, P. (2003). *Guide to workforce planning in local authorities, employers' organization for local government*, London.

Rosa, J. L. A., Ebecken, N. F., & Costa, M. C. A. (2003). Towards on an optimized parallel KNN-fuzzy classification approach. *WIT Transactions on Information and Communication Technologies*, 29.

Scott-Young, C., Samson, D. (2008). Project success and project team management: evidence from capital projects in the process industries. *J. Oper. Manag.* 26 (6), 749–766.

Song, Y., Huang, J., Zhou, D., Zha, H., & Giles, C. L. (2007, September). Iknn: Informative k-nearest neighbor pattern classification. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 248-264). Springer, Berlin, Heidelberg.

Tabassi, A.A., Ramli, M., Bakar, A.H.A., (2012). Effects of training and motivation practices on teamwork improvement and task efficiency: the case of construction firms. *Int. J. Proj. Manag.* 30 (2), 213–224

Tam, V. W., Tam, C. M., Chan, J. K., & Ng, W. C. (2006). Cutting construction wastes by prefabrication. *International Journal of Construction Management*, 6(1), 15-25.

Tang, J.C.S., Karasudhi, P. and Tachopiyahoon, P. (1990), “Thai construction industry: demand and projection”, *Construction Management and Economics*, Vol. 8, pp. 249-57.

Tixier, A. J. P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2016). Application of machine learning to construction injury prediction. *Automation in construction*, 69, 102-114.

- Tommelein, I. D., Levitt, R. E., & Hayes-Roth, B. (1992). Site-layout modeling: how can artificial intelligence help? *Journal of construction engineering and management*, 118(3), 594-611.
- Tsui, A.S. (1987). Defining the activities and effectiveness of the human resource department: a multiple constituency approach. *Hum. Resour. Manag.* 26 (1), 35–69
- Tsui, A.S., Milkovich, G.T. (1987). Personnel department activities: constituency perspectives and preferences. *Pers. Psychol.* 40 (3), 519–537
- Turner, J. R., & Müller, R. (2005). The project manager's leadership style as a success factor on projects: A literature review. *Project management journal*, 36(2), 49-61.
- Turner, P. (2002). *Strategic human resource forecasting*. London: CIPD.
- Wang, M., Cui, Y., Wang, X., Xiao, S., & Jiang, J. (2017). Machine learning for networking: Workflow, advances and opportunities. *Ieee Network*, 32(2), 92-99.
- Wang, Y., & Yao, Y. (1999). Sources of China's economic growth, 1952–99: incorporating human capital accumulation. The World Bank.
- Wauters, M., & Vanhoucke, M. (2017). A nearest neighbour extension to project duration forecasting with artificial intelligence. *European Journal of Operational Research*, 259(3), 1097-1111.
- Werther, W. B., & Davis, K. (1982). *Personnel management and human resources*. New York: McGraw-Hill.
- Widyanty, W., Daito, A., Riyanto, S., & Nusraningrum, D. (2020). Gaining a competitive advantage through strategic human resource management in Indonesian construction industry. *Management Science Letters*, 10(9), 2021-2028.
- Wild, A. (2002). The unmanageability of construction and the theoretical psycho-social dynamics of projects. *Engineering Construction and Architectural Management*, 9(4), 345-351.
- Witten IH, Frank E, Hall MA. (2011). *Data mining: practical machine learning tools and techniques*. Burlington, Massachusetts, USA: Morgan Kaufmann Publishers.
- Wong, J., Chan, A., & Chiang, Y. H. (2004). A critical review of forecasting models to predict manpower demand. *Construction Economics and Building*, 4(2), 43-56.

- Wong, J.M.W., Chan, A.P.C. and Chiang, Y.H. (2003), “Determinants of construction manpower demand: a review from literature and practitioners’ experience”, 2nd International Conference on Construction in the 21st Century (CITC-II), Hong Kong, 10-12 December 2003, pp. 158-63.
- Wright, P.M., Boswell, W.R. (2002). Desegregating HRM: a review and synthesis of micro and macro human resource management research. *J. Manag.* 28 (3), 247–276
- Wu, J. D., & Chan, J. J. (2009). Faulted gear identification of a rotating machinery based on wavelet transform and artificial neural network. *Expert Systems with Applications*, 36(5), 8862-8875.
- Xie, Q. (2020). Machine learning in human resource system of intelligent manufacturing industry. *Enterprise Information Systems*, 1-21.
- Yaseen, Z. M., Ali, Z. H., Salih, S. Q., & Al-Ansari, N. (2020). Prediction of risk delay in construction projects using a hybrid artificial intelligence model. *Sustainability*, 12(4), 1514.
- Zaremba, W., Sutskever, I., & Vinyals, O. (2014). Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.
- Zhou, X., Zhu, X., Dong, Z., & Guo, W. (2016). Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *The Crop Journal*, 4(3), 212-219.
- Zhu, X., Zhang, S., Jin, Z., Zhang, Z., & Xu, Z. (2010). Missing value estimation for mixed-attribute data sets. *IEEE Transactions on Knowledge and Data Engineering*, 23(1), 110-121.
- Zwikael, O., Unger-Aviram, E. (2010). HRM in project groups: the effect of project duration on team development effectiveness. *Int. J. Proj. Manag.* 28 (5), 413–421.

Appendix

The following figures elaborate some parts of the python code for developing the application:

```
import tkinter as ttk
from matplotlib.figure import Figure
import numpy as np
from matplotlib.backends.backend_tkagg import (FigureCanvasTkAgg, NavigationToolbar2Tk)
import seaborn as sns
import tensorflow as tf
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
from math import sqrt
from sklearn import preprocessing
from sklearn.metrics import r2_score
from sklearn.preprocessing import LabelEncoder
from sklearn.neural_network import MLPRegressor
from sklearn.metrics import mean_absolute_error
from sklearn.model_selection import KFold
from sklearn.preprocessing import StandardScaler
from sklearn.utils import resample
import pickle
import matplotlib.pyplot as plt
from scipy.stats.kde import gaussian_kde
import scipy.stats as stats
```

Figure A. Loading the Libraries

```
root = Tk()
root.title("Curve Forecasting Model")
# Add a grid
mainframe = Frame(root)
mainframe.grid(column=0, row=0, sticky=(N,W,E,S) )
mainframe.columnconfigure(0, weight = 1)
mainframe.rowconfigure(0, weight = 1)
mainframe.pack(pady = 100, padx = 100)
root.geometry("650x750")
```

Figure B. Setting up the Main Window of the Interface (size, title, etc.)

```

# Create a Tkinter variable
tkvar1 = StringVar(root)
tkvar2 = StringVar(root)
tkvar3 = StringVar(root)
#tkvar4 = StringVar(root)

# Dictionary with options
choices1 = { 'Concrete', 'Culvert', 'Cut & Fill',
             'Excavation', 'Piping',
             'Site Drainage', 'Steel'}
choices2 = { 'Carpenter', 'Welder',
             'Labor', 'Operator', 'Pipefitter'
            }
choices3 = {'High', 'Medium', 'Low'}
#choices4 = { 'Construction Management', 'Design Bid Build', 'Design Build', 'In-House'}

tkvar1.set('Choose') # set the default option
tkvar2.set('Choose')
tkvar3.set('Choose')
#tkvar4.set('Choose')

```

Figure C. Defining Dropdown Buttons and their Values

```

Label(mainframe, text="Enter the information :", font=('helvetica', 13, 'bold')).grid(row = 0, column = 3)

popupMenu = OptionMenu(mainframe, tkvar1, *choices1)
Label(mainframe, text="Work Package Category").grid(row = 3, column = 1)
popupMenu.grid(row = 4, column =1)

popupMenu = OptionMenu(mainframe, tkvar2, *choices2)
Label(mainframe, text="Resource Type").grid(row =3, column = 3)
popupMenu.grid(row = 4, column =3)

popupMenu = OptionMenu(mainframe, tkvar3, *choices3)
Label(mainframe, text="Duration").grid(row =3, column = 5)
popupMenu.grid(row = 4, column =5)

```

Figure D. Defining the Labels and Texts - Placing the Buttons and their Labels


```

if tkvar3.get() == 'Very High':
    LowLimit = 25
    HighLimit = 200
if tkvar3.get() == 'High':
    LowLimit = 15
    HighLimit = 25
if tkvar3.get() == 'Medium':
    LowLimit = 8
    HighLimit = 15
if tkvar3.get() == 'Low':
    LowLimit = 2
    HighLimit = 8

```

Figure E. Defining the Different Categories of Duration Based on the Result of Clustering

```

data1 = data[(data['WP_Type'] == tkvar1.get())]
data1 = data1[(data1['Resource_Type'] == tkvar2.get())]

```

Figure F. Filtering the Records Based on the Selected Attributes by User

```

data1=data1.loc[:,['Total_Hours','Duration','TU','Hours']]
data2 = data1[(data1['Duration'] > LowLimit) & (data1['Duration'] < HighLimit)]

```

Figure G. Dropping Additional Attributes and Filtering Records Based on Selected Duration

```

#data2 = data[(data['WP_Type'] == tkvar1.get())]
duration = data2["Duration"].mean()
duration = int(duration)
print(duration)

```

Figure H. Calculating the Duration

```

cv = KFold(n_splits=8, random_state=None, shuffle=False)
for train_index, test_index in cv.split(X):
    X_train, X_test, y_train, y_test = X[train_index], X[test_index], y[train_index], y[test_index]
    RNN.fit(X_train, y_train)
    pred_test_mlp = RNN.predict(X_test)
    MAE.append(mean_absolute_error(y_test, pred_test_mlp))

TotalError = TotalError + float(np.mean(MAE))

```

Figure I. k-fold Cross Validation for Training and Testing – Calculating Error Value

```

canvas2 = ttk.Canvas(root, width = 600, height = 60, relief = 'raised')
canvas2.pack()
label4 = ttk.Label(root, text="The Mean Absolute Error (MAE) is : " + str(round(TotalError,1)) + " Labor-Hours" ,
                    font=('helvetica', 11, 'bold'))
canvas2.create_window(300, 40, window=label4)

label5 = ttk.Label(root, text="The Total Hours is : " + str(round(TotalHours,1)) + " Labor-Hours" ,
                    font=('helvetica', 11, 'bold'))
canvas2.create_window(300, 15, window=label5)

```

Figure J. Setting up the Canvas for the Predicted “Total Hour” and “Error”

```

fig = Figure(figsize = (12, 8), dpi = 100)
x=np.arange(1,int(duration)+1,1)
y=PredictedHours
plot1 = fig.add_subplot(111)
plot1.plot(x,y , marker='X', label='Predictions', color='blue')
plot1.legend(loc="upper right")
plot1.set_xlabel('Time (Day)')
plot1.set_ylabel('Labor-Hours')
plot1.LineWidth = 4
canvas = FigureCanvasTkAgg(fig, master = root)
canvas.draw()
canvas.get_tk_widget().pack()

```

Figure K. Plot Configuration and Setting up a Canvas to Elaborate the Plot