

**MACHINE LEARNING FOR DATA CLASSIFICATION IN CONSTRUCTION**

**PROJECT PLANNING**

**BY**

**LAU ZHENG LIANG**

**A REPORT**

**SUBMITTED TO**

**Universiti Tunku Abdul Rahman**

**in partial fulfillment of the requirements**

**for the degree of**

**BACHELOR OF INFORMATION SYSTEMS (HONOURS) BUSINESS**

**INFORMATION SYSTEMS**

**Faculty of Information and Communication Technology**

**(Kampar Campus)**

**JUNE 2023**

**MACHINE LEARNING FOR DATA CLASSIFICATION IN  
CONSTRUCTION PROJECT PLANNING**

**BY**

**LAU ZHENG LIANG**

**A REPORT**

**SUBMITTED TO**

**Universiti Tunku Abdul Rahman**

**in partial fulfillment of the requirements**

**for the degree of**

**BACHELOR OF INFORMATION SYSTEMS (HONOURS) BUSINESS**

**INFORMATION SYSTEMS**

**Faculty of Information and Communication Technology**

**(Kampar Campus)**

**JUNE 2023**

## REPORT STATUS DECLARATION FORM

**Title:** MACHINE LEARNING  
FOR DATA CLASSIFICATION  
IN CONSTRUCTION PROJECT PLANNING

**Academic Session:** Jun 2023

I LAU ZHENG LIANG  
**(CAPITAL LETTER)**

declare that I allow this Final Year Project Report to be kept in  
Universiti Tunku Abdul Rahman Library subject to the regulations as follows:

1. The dissertation is a property of the Library.
2. The Library is allowed to make copies of this dissertation for academic purposes.

Verified by,



(Author's signature)



(Supervisor's signature)

**Address:**

A2-02-17, Jalan Indah 23/1,  
Taman Bukit Indah 2,  
81200 Johor Bahru, Johor.

Cik Zanariah Binti Zainudin  
Supervisor's name

**Date:** 7 September 2023

**Date:** 8 September 2023



## DECLARATION OF ORIGINALITY

I declare that this report entitled “**MACHINE LEARNING FOR DATA CLASSIFICATION IN CONSTRUCTION PROJECT PLANNING**” is my own work except as cited in the references. The report has not been accepted for any degree and is not being submitted concurrently in candidature for any degree or other award.

Signature :  \_\_\_\_\_

Name : Lau Zheng Liang

Date : 7 September 2023

## **ACKNOWLEDGEMENTS**

I would like to special thanks to my supervisor, Miss Zanariah Binti Zainudin who have guild me to explore my idea of the data mining and machine learning field study. This project is learn develop a Machine Learning into the real business with my idea. Additionally, she is kind to guild me when I was faced on the problems. I wish this project will lead me gain the knowledge from machine learning in the real world. I honest to glad she was my supervisor of my final year project.

Other side, I want to give a thousand thanks to my parents who support me my education. They fully support my talent to study this programme which is business of information system. I will always give you my whole loves and respects in the future. I always keep remind my parents said, “How much you hard work, how much you will deserved”.

## **ABSTRACT**

Machine Learning is a tool of Artificial Intelligence to use the algorithms and statistical models to enable a machine to learn the concept and make predictions or decision as an outcome. The Machine learning is a tool created from statistical principles and expanded into the world. The concept of the Machine Learning is the ability of the machine able to learn the situation with algorithms rules and make a predictions or decision. Hence, it is useful to develop in different area in the real world. The Machine Learning can develop into many areas which are Gaming, Data Mining and Analysis, Recommendation System, Financial Management and so on.

In the country of Malaysia, most of the project companies are process the traditional method to analysis the project budget. In beside of the situation, most of the companies fear the Artificial Intelligence is taking over on the business. As known, the project budget is important to a project company to make sure the project they taken is reasonable and suitable from the client given. The project company choose the project they take and discuss with the client on the project's budget. Most of the reason the project taken was over the budget is lack of understand of the project and spend a lot of cost on time management and project scope.

In the situation, Machine learning model helps the project company to analyze the project and decision the advice in project planning management. It helps to most project companies increase business strategy with Machine Learning model. Resulting the project company easier make decision in the project planning without distressed. Based on business information management, Machine Learning is the most recommended technologies tool to supports many project companies and SME in their business growing. This project is to proof the Machine Learning can develop for the local business project company.

# TABLE OF CONTENTS

|  |            |
|--|------------|
| <b>TITLE PAGE</b>  | <b>i</b>   |
| <b>REPORT STATUS DECLARATION FORM</b>  | <b>ii</b>  |
| <b>FYP THESIS SUBMISSION FORM</b>  | <b>iii</b> |
| <b>DECLARATION OF ORIGINALITY</b>  | <b>iv</b>  |
| <b>ACKNOWLEDGEMENTS</b>  | <b>v</b>   |
| <b>ABSTRACT</b>  | <b>vi</b>  |
| <b>TABLE OF CONTENTS</b>   | <b>vii</b> |
| <b>LIST OF FIGURES</b>   | <b>x</b>   |
| <b>LIST OF TABLES</b>  | <b>xi</b>  |
| <b>LIST OF ABBREVIATIONS</b>   | <b>xii</b> |
| <br>   |            |
| <b>CHAPTER 1 INTRODUCTION</b>  | <b>1</b>   |
| 1.1 Problem Statement and Motivation   | 1          |
| 1.2 Research Objectives  | 3          |
| 1.3 Research Scope   | 5          |
| 1.4 Research Contributions   | 6          |
| 1.5 Report Organization  | 7          |
| <br>   |            |
| <b>CHAPTER 2 LITERATURE REVIEW</b>   | <b>8</b>   |
| 2.1 Introduce Technologies Used  | 8          |
| 2.1.1 Hardware   | 8          |
| 2.1.2 Software Tools   | 9          |
| 2.2 Introduce the Previous Works of Classification in Project Management                   | 11         |
| 2.2.1 To Develop a Machine Learning Model to Predict Cost Overruns in Construction Project | 13         |
| 2.2.2 To Develop Machine Learning for Budget Forecast in Capital Construction              | 14         |
| 2.2.3 To Develop Machine Learning for Project Budget Prediction                            | 15         |



|   |           |
|---|-----------|
| 2.2.4 To Determinants the Successful Budgeting with<br>Machine Learning Models    | 16        |
| 2.2.5 Determine the Level of Bid Price from Risk Factors                          | 18        |
| 2.2.6 Classification for Project Feature with Machine<br>Learning Model           | 19        |
| 2.2.7 To Fine-Tuning the Parameters of the Algorithms                             | 20        |
| 2.3 Summary of Authors' Previous Works of Classification in<br>Project Management | 21        |
| <b>CHAPTER 3 RESEARCH MODEL</b>   | <b>23</b> |
| 3.1 Research Methodology  | 23        |
| 3.2 Summary on Research Model   | 28        |
| <b>CHAPTER 4 DATA UNDERSTANDING</b>   | <b>29</b> |
| 4.1 Introduction of Dataset   | 29        |
| 4.2 Data Features Description   | 30        |
| 4.3 Data Visualization  | 33        |
| 4.4 Summary of Data Understanding   | 36        |
| <b>CHAPTER 5 DATA PREPARATION</b>   | <b>37</b> |
| 5.1 Data Preprocessing  | 37        |
| 5.1.1 Data Quality Problem  | 37        |
| 5.1.2 Data Cleaning   | 38        |
| 5.2 Split Data Training and Test Model  | 43        |
| <b>CHAPTER 6 ALGORITHM SELECTION AND PERFORMANCE<br/>EVALUATION</b>               | <b>44</b> |
| 6.1 Classification Algorithm Selection  | 44        |
| 6.2 Modelling   | 44        |
| 6.3 Test Set Model  | 47        |
| 6.4 Summary of Postprocessing   | 52        |

|   |           |
|---|-----------|
| <b>CHAPTER 7 CONCLUSION AND RECOMMENDATIONS</b> | <b>53</b> |
| 7.1 Conclusion                                  | 53        |
| 7.2 Recommendations                             | 55        |
| <b>REFERENCES</b>                               | <b>56</b> |
| <b>WEEKLY LOG</b>                               | <b>59</b> |
| <b>POSTER</b>                                   | <b>65</b> |
| <b>PLAGIARISM CHECK RESULT</b>                  | <b>66</b> |
| <b>FYP2 CHECKLIST</b>                           | <b>73</b> |

## LIST OF FIGURES

| <b>Figure Number</b> | <b>Title</b>                                   | <b>Page</b> |
|----------------------|--|-------------|
| Figure 3.1           | Research Procedure                             | 23          |
| Figure 4.1           | Project Geographic District in project records | 33          |
| Figure 4.2           | Project Type of project records                | 34          |
| Figure 4.3           | Project Phase Name                             | 35          |
| Figure 5.1           | Flowchart of the data cleaning                 | 38          |
| Figure 5.2           | Feature Selection with Correlation Matrix      | 40          |
| Figure 5.3           | Correlation Matrix after Feature Selection     | 41          |
| Figure 6.1           | Confusion Matrix of Logistic Regression        | 48          |
| Figure 6.2           | Confusion Matrix of Decision Tree              | 49          |
| Figure 6.3           | Confusion Matrix of Random Forest              | 50          |
| Figure 6.4           | Confusion Matrix of Logistic Regression        | 51          |

## LIST OF TABLES

| <b>Table Number</b> | <b>Title</b>                            | <b>Page</b> |
|---------------------|---|-------------|
| Table 2.1           | Specifications of laptop                | 8           |
| Table 2.2           | Table of Previous Works                 | 12          |
| Table 6.1           | Accuracy Performance of Each Algorithms | 47          |
| Table 6.2           | Score Matrix of Logistic Regression     | 48          |
| Table 6.3           | Score Matrix of Decision Tree           | 49          |
| Table 6.4           | Score Matrix of Random Forest           | 50          |
| Table 6.5           | Score Matrix of SVM                     | 51          |

## **LIST OF ABBREVIATIONS**

|            |                                    |
|------------|------------------------------------|
| <i>KNN</i> | K-Nearest Neighbors Algorithm      |
| <i>ANN</i> | Artificial Neural Networks         |
| <i>SVM</i> | Support Vector Machines            |
| <i>SME</i> | Small and Medium-sized Enterprises |
| <i>B2B</i> | Business-to-Business               |

# Chapter 1

## Introduction

In this chapter, the research of the background and motivation to develop the Machine Learning model for a construction project company is presented in this chapter. This is a project research based to review and develop the Machine Learning classification algorithms to achieve the project scope for construction project profitable analysis. To achieve the research scope, the objectives must be created and followed until the goal is achieved.

### 1.1 Problem Statement and Motivation

In the real world, Machine Learning model helps a company business in understanding their growth strategies. The problem of the business growing is lack of understand and decision making in the project planning between client expectation and actual project cost spending, and this result is lost profits. To understand the business growing, a huge dataset is a main major to collect and record the projects taken and carried out, and developed a Machine Learning to analyze the profit earned.

There are problem statements the construction company faced is spend a lot of time and the cost of the budget to process a construction project planning in decision making. It is difficult to deliver projects to clients on time when decision-making in project management is lacking. The reason is the human decision is slower and much of human error in the project planning budget phases. It is difficult to measure the actual project cost budget is higher than the expected cost. Furthermore, the construction project company has receive a different type of project and project phases. Each of the project categories doesn't have classify into multiclass, the different types of projects must have a standard budget to ensure the company won't lose the profile from the project planning. [1]

The last of the problem statements is there are many construction project companies are going use the traditional method to plan the project budget with their client such as meeting and interviewing. This approach has an impact on businesses that spend a lot of money on projects that fall short of client expectations or change because of unclear ideas. Work hard doesn't mean the employees use the human power to force on the old methods, the method must keep updating and renewing to reduce the human power and increase work infective.

In the motivation towards to research project goal is aimed to develop a Machine Learning model to classification from the project planning and budget dataset. The goal of developing a Machine Learning model for a construction project company is used to improve project planning efficiency and reduce costs during the planning stages [2]. To do so, this project musts to assess and explore Machine Learning classification algorithms to enable to choose the most suitable to develop with the huge dataset of the construction project.

## **1.2 Research Objectives**

To complete the scope of the research, the objectives must be required to achieve the research of develop Machine Learning classification model for the construction project company which relate in real business. These are the objectives to ensure completed deliver to the research goal.

### **1.2.1 Find a Dataset related with Construction Projects**

A dataset is required to the research project with the goal to develop a Machine Learning model of classification algorithms as a case study. The research project takes a dataset and spilt into a training and test set for the machine to learning with algorithm rules and tuning parameter of the algorithms. Dataset must select related with construction projects from the internet source to suitable achieve the research scope and goal. After selected the dataset, the data features must be understand the information presented and make into the visualization which is contain in dataset.

### **1.2.2 To Enhance Data Preprocessing for Improved Analysis and Insights**

The research project must deal with a quality issue in a dataset that was chosen from an internet source. The dataset's information and description detail from internet source must understand before process to identify the quality problem. After identifying the dataset quality problem, the data cleaning methods have to selected to process into cleaned dataset to improve a machine to analysis and maintain the algorithms of the accurate performance. The preprocessing has taken longer times to deal the dataset quality problem on this research project.

### **1.2.3 To Classify the Project Remain Budget**

This is the main research objective to classify for the construction project planning management. The research scope is enable the project manager to use the Machine Learning model to classify the project profitable. The remain budget the company earn from the client, it became a company profitable. Besides, it can increase the efficiency of the budget planning management. Understand the construction dataset's data feature of the relationship which serious effect to the company profitable. This research project of the primary goal accomplishes in order to fulfill the project's scope.



#### **1.2.4 To use Fine-Tuning for Optimal Parameters on Classification Algorithms**

Fine tuning the parameters on each classification algorithm is important to avoid a machine overfitting and underfitting happened. In the same case, the hyper tuning is to prevent a machine has higher accuracy learning with the regular algorithms. The parameters of the algorithm are manual handling and ensure the machine is not overfitting or underfitting in the test model. Those algorithms' accuracy of the parameters handling must over the 80% to ensure a machine has higher accuracy to learning and selected the higher performance to analysis the algorithm selected.

#### **1.2.5 To get Best Performance of Classification Algorithm**

In the test model, the higher performance of the classification algorithm has to selected to develop to the project planning management. The analysis of the test model is an important phase to analysis the accuracy the actual accuracy from the test model to ensure it has not overfitting or underfitting case happened. Even though it has already occurred, the research methodology must return to modeling session to fine-tune the algorithmic parameters and avoid a serious situation. After the case, the research project must select a classification algorithm with higher performance to analyze the development that is reasonable for the construction project company.

### **1.3 Research Scope**

The research of the scope is allow the construction company to use the Machine Learning model to analyze client expectations on project expenses and actual project spending costs in project planning. The Machine Learning model with classification algorithms must deliver a result to the construction project manager for easier analysis in planning. The algorithms used in this research project are Logistic Regression, Decision Tree, Random Forest, and *SVM*. The project scope is help a project more productivity and efficiency in project management in business growing. As a way to prevent project planning mistakes, the workplace must implement modern techniques as well as the employee has to participate effectively. Faster deliver on the decision-making process is the way to full fit the company growing in the business.

To deliver the research scope, the dataset must related the construction project, which is contain the project information, project phases, estimate cost spending and actual spend budget. The dataset must be research to achieve the research objective (Section 1.2). The data understanding is important to understand the data variables before process the data cleaning in preprocessing. Furthermore, the machine learning algorithms must be created and train from the training set. The more algorithms created and measured the performance with the test set. In the case, the higher accuracy of the algorithms will be selected and summary the algorithms used in the construction project company

#### 1.4 Research Contribution

This research of the contribution is increase the business strategic that each of the project budget contain standard budget with category. This project is to improve decision management of the decision-making effectiveness. The reason of develops a Machine Learning classification model is to avoid the human mistake to make decision on the project in the planning phases. The company wants to avoid costly overspending and developing overtime. This project will helps a company going towards into the modern technology in the project planning. To do that, the Machine Learning must be develop for classifier the dataset of project planning recorded to analysis. [3]

To develop a Machine Learning of data classification towards the construction project company, the first aim is research and review the Machine Learning classification algorithms from the authors' previous works. The company is planning for each of the project must have a standard budget with a category. Classification is suitable to use for this project for company business used. The classification algorithms have contain many models which are Logistic Regression, Decision Tree, *KNN*, *ANN*, Random Forest, *SVM*, and Naive Bayes. Each classification methods should be research to understand the logistic function usable to deal the dataset into result for project planning used. After research the knowledge and previous work, the research will identify the company's dataset that needs to be analyzed by using classification techniques. After that, it selects the classification algorithm to suitable used for the classification projects for the project planning.

This project is to attract the construction project company use modern technology more than traditional human power. Nowadays, the local country of the *SME* and companies are going digital business use the technology to operate the business into globalization. This objective is wished to develop a Machine Learning for the project company to improve productivity without human power mistake [4]. The technology strategic helps company increase the productivity and effectivity to save the costs on the resources. This objective is a study case that use the information technology system to suitable use in the business analysis to attract project company to analysis in every project they receive.

## **1.5 Report Organization**

This report has been organized into 6 chapters, which are Chapter 1 Introduction, Chapter 2 Literature Review, Chapter 3 Research Methodology, Chapter 4 Data Understanding, Chapter 5 Data Preparation, Chapter 6 Algorithm Selection and Performance Evaluation, and Chapter 7 Conclusion and Recommendations. The first chapter is introduced the research of the scope and objectives to achieve the goals in the construction project planning. The second chapter is introduced the technologies used in this project, which are hardware and software. Furthermore, the chapter reviewed and summary on the previous works to understand the articles developed the Machine Learning model used to solve the problem in project management. The third chapter is introduced the research methodology used to develop a Machine Learning model in this project. Chapter fourth is summary the information of dataset used in this research project. Fifth chapter is summary the dataset quality issue and introduce the data cleaning methods in preprocessing. Sixth chapter is introduce the data modelling in algorithms model and analyze the performance of algorithms. The last chapter is conclusion the compare of the research project, and summary the limitations and improvements of the research project.

# Chapter 2

## Literature Review

This chapter is introduced the technologies used in this project, which are hardware, and software. The software provided information about how to develop a Machine Learning model. The chapter also review and summary the authors of the previous works to understand the Machine Learning classification algorithms used to face the situation into the real business world.

### 2.1 Introduce Technologies Used

This section is to introduce the technologies used in this project which are hardware and software tools. These technologies help the project processed develop a machine learning model towards the project objectives and goal. The hardware and software tools which are used in this project are introduced in these subsections (Section 2.1.1 - 2.1.2).

#### 2.1.1 Hardware

The hardware device is used to processed develop a Machine Learning model. It used to provide the data visualization, data preprocessing, modelling, and analysis. The hardware used to analyze the results of data mining in order to understand and make decisions based on project budget planning. The hardware device's specifications are listed in this below:

| Description      | Specifications   |
|------------------|--|
| Model            | Acer Nitro 5 2022 (AN515-58-72ND)                        |
| Processor        | 12 <sup>th</sup> Gen Intel(R) Core(TM) i7-12700H 2.70GHz |
| Operating System | Microsoft Window 11 Home                                 |
| Graphic Card     | Nvidia GeForce RTX3060 6GB GDDR6                         |
| Memory           | 16GB DDR4 3200Mhz  |
| Storage          | 512GB PCIe NVMe SSD                                      |

*Table 2.1 Specifications of laptop*

## **2.1.2 Software Tools**

These are the software tools are required process to develop a Machine Learning model. Each of the tool is used to processed data mining the dataset and select the suitable classification algorithm to classify the data into outcome. Hereafter, it will picked the higher accuracy of the classification algorithm and used in the project planning management. These tools are introduce which used in the research project (Section 2.1.2.1 – 2.1.2.4).

### **2.1.2.1 Microsoft Excel**

In this project, Microsoft Excel is required used to collect a dataset and access to the machine to learning the concept. The dataset named is “Capital Project Schedules and Budgets” which contain the construction project planning recorded. In the dataset, there are many different types of the project, and each of the project has different time schedule phase. Some of the projects are in processing and some of the project is done. More, some of the actual budget projects are overweight with the planning cost budget. Therefore, the dataset is important for the Machine Learning to process decide a project budget which is acceptable or unacceptable.

### **2.1.2.2 Anaconda**

Anaconda is an open-source platform allow to write and execute code with the Python language. Anaconda has provide the data science tools used for the data science needed to development a Machine Learning model and deep learning model. It can provide the package and framework to work for analysis in current environment. It is unfamiliar with the Visual Studio Code which must install a lot of packages to allow development the Machine Learning model. Aside, Anaconda has many tools of data science to do the data analysis in Python language. it automatic installed.

### **2.1.2.3 Jupyter Notebook**

Jupyter Notebook is a software allows analysis and development a Machine Learning model in visualization. The Jupyter Notebook provides the Python language to develop the Machine Learning to processed data analysis with the dataset. It performs data visualization, clustering, classification, and regression to do data mining in this software. It is arranged step by step to develop a code, visual, and accuracy outcome. Most of the data analyst and data science are using this software to do the data analysis for the company to identify market target for the business growth.

### **2.1.2.4 Python**

Python language is used in the Jupyter Notebook. Python provides libraries used to data mining with algorithms rules. The Python libraries used are Matplotlib, NumPy, Pandas and Scikit-Learn. These are the powerful libraries used to process data mining. These are commonly used for Machine Learning to classification and regression with the dataset. Matplotlib is used to develop a data visualization to understand the concept of the dataset. NumPy stands for Numerical Python. It is a library tool that provides numerical computation use for array and matrix data including the logical, random math, shape array. Pandas is a library tool to read the dataset and do the data frames. The Pandas can do the task to work on the data include data cleansing, merges and join, and index the dictionary. Scikit-Learn is a useful library use to identifying and categorizing the data based on patterns in classification. This library allows to measure the classification calculation to process the classifier function.

## 2.2 Introduce the Previous Works of Classification in Project Management

These are previous works which were studied with the objective to comprehend these authors used the Machine Learning Classification Model to analyze the data on project risk and effectively plan projects in the construction project company. This sub-chapter is to understand the authors problem solving of developed Machine Learning in the business. These are the previous works have been research in latest years (2020 - 2023) to ensure the machine learning technologies must be similar with the real-world generation as a study case. Afterward, these previous works are summaries in the sub-topic (Section 2.2.1 – 2.2.7).

| No. | Authors                                | Title  | Dataset                                       | Algorithms Used                                     | Result   |
|-----|--|--|---|---|--|
| 1   | (T. Aung, Sui Reng Liana et al., 2023) | Using Machine Learning to Predict Cost Overruns in Construction Project [5]              | Construction Projects from various sources    | Linear Regression<br><i>SVM</i><br><i>ANN</i>       | Linear Regression is a higher MAE and RMSE, which are 0.725 and 0.938. |
| 2   | (P. Ghimire, S. Pokharel et al., 2023) | Machine Learning-based Prediction Models for Budget Forecast in Capital Construction [6] | Utilizes data from the NYC Open Data database | Decision Tree<br>Random Forest<br>Linear Regression | Decision Tree is a higher performance in 96% of the total variability. |



|   |   |  |  |   |  |
|---|---|--|--|---|--|
| 3 | (W. Kusunghum, K. Srinavin et al., 2022)  | Government Construction Project Budget Prediction Using Machine Learning [7]   | Thailand Government Procurement electronic system (e-GP) | <i>KNN</i>  | Accuracy – 0.80<br><br>Precision:<br>Yes case = 0.56<br>No case = 0.86                 |
| 4 | (M. Kunnathuvalappil, 2021)               | Investigating the Determinants of Successful Budgeting with SVM and Binary models [8]  | Project from authors' company                            | <i>SVM</i>  | Coordination – 0.42<br>Control – 0.36<br>Coordination – 0.058<br>Participation – 0.058 |
| 5 | (YeEun Jang, Heong Wook Son et al., 2021) | Classifying the Level of Bid Price Volatility Based on Machine Learning with Parameters from Bid Documents as Risk Factors [9] | California Department of Transportation (Caltrans), US   | <i>ANN</i><br>Decision Tree<br><i>SVM</i><br><i>KNN</i> | <i>ANN</i> is higher performance in model 1 and model 2.                               |
| 6 | (Ching Ling Fan, 2020)                    | Evaluation of Classification for Project Feature with Machine Learning Algorithms [10]   | Public Works Bid Management System (PWBMS)               | Decision Tree<br><i>ANN</i><br><i>SVM</i>               | Not Mentioned.   |

Table 2.2 Table of Previous Works

### **2.2.1 To Develop a Machine Learning Model to Predict Cost Overruns in Construction Project**

In [5], the objective of the previous work is the authors wish the business company enable would use the Machine Learning to predict the project cost and replace the traditional cost estimation. The authors noticed that expert judgment and parametric estimation, these two conventional methods used in construction project management may not provide a true estimate of the project cost. The result in a project had inaccurate cost prediction and increased risk of overruns.

The authors used construction projects dataset from various sources to develop a machine learning model to predict the project cost. In the [5] dataset used, it contains 250 construction projects with different project type, which are residential, commercial, infrastructure, and industrial construction projects. In each of the projects recorded have information including project size, labor costs, materials costs, and initial estimated costs. Plus, the dataset has been includes the actual costs and resulting cost overruns for each project. Therefore, the authors decide used Linear Regression, *SVM* and *ANN* to develop for the work and tried use algorithms to take over the traditional method.

In [5] modelling, the authors used mean absolute error (*MAE*) and root mean square error (*RMSE*) to evaluate the performance of the machine learning model to ensure the algorithms is in higher accuracy in between expected and actual cost of project overruns. In results, the authors found out that the higher performance of the algorithm is Linear Regression, and it is over the traditional method which are expert judgment and parametric estimation, with a 41.24% in *MAE* and 40.63% in *RMSE*. These are affect from the most project parameters to predict the cost overruns which is initial estimated costs. In the end, the Linear Regression selected to develop a machine learning to predict the project cost in the construction project company.

### 2.2.2 To Develop Machine Learning for Budget Forecast in Capital Construction

In [6] the authors' idea of the previous work is the project cost allocation often prevents the organization from controlled and the resulting in an unreliable forecast of the project profit or losses after the construction is completed. From the idea, the authors wished to build a professional work to analyze the project cost in the project lifecycle such as machine learning model. The authors wished to increase the efficiency of work environment in used machine learning model and prevent the cost is overbudget and project schedule delays.

In [6] data preparation, the authors take a dataset from the open-source dataset updated in January 2023, which is NYC Open Data database. The dataset has 3157 projects recorded and 16 variables for capital projects dating from 1995 to 2023. These projects are from the State of New York, USA. Each of the project has five variable which are project category, project phase, budget forecast, total schedule changes and total budget changes. As understand the dataset, the authors selected decision tree, random forest, and linear regression to develop a machine learning and selected the higher performance to the analysis of the predict cost. After selecting the algorithm, the authors select coefficient of determination ( $R^2$ ), mean absolute error (*MAE*), and root mean squared error (*RMSE*) to measure the algorithm's accuracy performance in this work.

Resulting of [6], the decision tree is a higher performance model and the most accurate prediction the project costs. The decision tree has a higher performance square value is 96% with 17.55 in *MAE* and 43.82 in *RMSE*. In the case, the authors mentioned the resulting the decision tree is the most accurate to predict in the nonlinearities between variables than other algorithms model. This model can used in company internal and external relationships to analyze the predict cost before starting the project lifecycle.

### 2.2.3 To Develop Machine Learning for Project Budget Prediction

In [7] previous work, the authors understand the big data used to develop a machine learning in the AI field. In the research, the authors found out Thai government used a big data to organization and departments technologies in project management. However, the problem they faced is lack of used this kind of technologies to develop in the works such as machine learning model. This situation causes a project management has lack of estimate cost and a high chance to lose the profit. This is a problem had happened in project cost management. To deal with this, the authors developed a machine learning model to analyze the project to the over funding profit as archive their objectives.

The authors [7] used Government Procurement electronic system (e-GP) dataset as a data preparation. This dataset contains project costs estimates and actual spending. This is a big data of the project the government planned and organization which authors mentioned. The attribute of the project has four types, which are location, department, type of project and procurement method. These are the types are affect the project cost in the project cost management. Afterward, the authors selected KNN methods to classify the type of project and analyze the outcome of the project cost.

In the result of [7], the authors found out the location is the most input attributes affect the project cost, and the higher project amount of method of the procurement is 601. Furthermore, most of the project type the government received is road project. The important is most of the projects are under budget, which mean the government was loss profit from the budget and haven't realized. Short summary, the authors used *KNN* methods to classify the project spending budget as a main objective in the project cost management. After developed, the model test is higher accuracy which is 0.86. In the model test, the higher precision of the outcome is No case which 0.86 in the project recorded. It means the project is under funding, the probability of the model analyzes the project as an over fitting is 0.56.

#### **2.2.4 To Determinants the Successful Budgeting with Machine Learning Models**

In [8] reviewed, the authors researched the types of determinants are affect to successful reach the project of budget. The types of determinants are referred as the kind of motivation or action is affected to the project performance and quality. The authors were realized most of the project manager just keep focus on the process track of the project and it affect the project quality is non-profit. The authors used a Machine Learning to understand most of the determinant is affected the success of the project budget. Therefore, the project managers should concentrate on the determinants which have a higher probability chosen in the analysis's outcome.

The dataset of the [8] previous work, the authors collected 470 projects from the company to understand the determinants are afford in the project management. In the dataset of the projects, it contains five features which are coordination, participation, budget control, communication, and motivation. These are the five features are affected the project budget which the project manager used the determinants in the project management. After that, the authors use *SVM* and Binary model to analysis the higher probability of the determinants of the outcomes which find out most determinant is affected on the project budget.

In the [8] modelling, the authors used *SVM* to classify the multiclass to measure the nearest distance of the class to avoid the overfitting. Next, the authors used Binary Model to measure the probability of observation on the determinants. The dataset was contain the dummy variable and noises. This is affect the machine difficult analysis the dataset from the noises. The authors changed the dummy variable into the numeric as refer from the dummy variable. Therefore, the authors easier used the Binary Model to measure the probability of the types of the determinants.

In [8] analyze, the total of the correlation coefficients for the determinants, which are control, coordination, motivation, communication, and participation. There are 0.73, 0.36, 0.058, 0.42 and 0.058. These are the probability of the determinants to affect the project reach the success and balance budget. In the correlation coefficients of the determinants, the higher probability of the determinants is the control. It means a project manager must control well on the project planning. A good project manager has to keep control on the project's timeline and budget spending. The project will profitable and best quality.

### 2.2.5 Determine the Level of Bid Price from Risk Factors

The construction project budget has calculated included different factor which are employees injury, environment, risk factors and so on. In [9] previous work, the authors found that the competitors bid the price on the different level of construction project in business project. As a knowledge, this is a *B2B* business that the construction company bid the project price to the client with another competitors. Most of the project is difficult to determine the bid price on the construction project based on the project behind risks factors. The authors wished developed a Machine Learning to determine the level of bid prices based on construction project of risk factors. Hence, this objective is ensure the bid price in contracts of construction project in fairness and earn profit from the competitor bid.

The authors [9] used a dataset from the California Department of Transportation (Caltrans) in the US as developed a Machine Learning. The dataset is contain 269 document which recorded the competitors bid succussed prices on a construction project. The construction project has contain working days, location, size, bid days, number of bidders, project type and quality. In the same case, the bid price of the construction project has included the risk factors which are time, cost, and quality. The problem of the dataset is there contain dirty data that the project data is recorded in unstructured text format, resulting the authors difficult quantitative analysis from the dataset.

As the result of postprocessing, the authors found that the *ANN* is the good performance of bid average risk from the training model and test model, which are 37.5% and 63.9%. In the same case, the accuracy of the rise algorithms the authors used are more than 58%. In the conclusion [9], the authors determine the *ANN* model is a good performance algorithm in the work and develop to the bid price to the business.

### **2.2.6 Classification for Project Feature with Machine Learning Model**

In this previous work [10] reviewed, the author used the classification algorithms to analysis the project quality and performance. The author found out the project feature is affected the project cost budget. It is mean the project feature is bad and affect the project process, the cost budget increase. The author faced the problem on the project management is there are a lot of projects are poor construction performance. In this situation, the author decided to use the machine learning model to classification the project of the feature to evaluation and avoid the project bad quality. The author wished to develop a Machine Learning model to classify the project's feature effected the planning and analysis the outcome to reduce the overweight costs and times.

In the [10],the dataset of the construction data selected from the Public Works Bid Management System (PWBMS) from 1993 to 2020 year. The dataset is contain the project construction which recorded by Taiwan's government Units. There are recorded 1015 projects and data features of the project are defect types, engineering levels, project costs, and construction progress. Among the data features, the projects decisions are classified into 4 types which are construction management, work quality, project program and design.

In the [10] preprocessing, the author was clustering the samples variable into the categories to ensure there is higher accuracy. The author used the sampling model to ensure the training set is reduces the overfitting. The project of the engineering levels feature is grouped by cluster analysis. The feature was clustered into six projects level. Unfortunately, there are two projects level are in the same grades. The author cluster into four project level that make sure the cluster is higher accuracy.

After the preprocessing, the authors planned select the algorithms which are decision tree, *ANN* and *SVM* to classification from the dataset learning. The authors wished to ensure a machine classification the meaningful outcome in higher accuracy from the project level. Following the author selected, the decision tree used to understand the variable which the machine used the rules to classify. Forward, the *SVM* helps analyze into higher accuracy of the multiclass project level. The *ANN* is analyze the new project of level in higher accuracy from the rule of decision tree and *SVM*.



### **2.2.7 To Fine-Tuning the Parameters of the Algorithms**

In [5], and [6] previous works research, the authors are use fine tuning the parameters of the algorithms used in their previous works. These are research project used fine-tuning on parameters in algorithms to prevent the overfitting or underfitting case in the modeling with the objective. Reasoning the algorithms the authors used to search a best parameter with the manual and develop the higher performance of algorithms to process postprocessing stage.

In [5], the authors researched the other previous work as the literature review to understand the purpose used the fine-tuning to increase the accurate performance of the algorithms. The authors found out the reasoning of used fine-tuning in parameter of the algorithms in their previous work [5] is reduce the financial risk and ensure the machine focus predict on cost overruns in the construction project. Therefore, the authors develop the algorithms they selected and provide fine-tuning to prevent noise on the machine.

According [6] previous work, the authors were fine-tuning the best parameters in algorithms used to improve the performance with cross-validation. The authors were mentioned reasoning used optimal parameters to prevent the machine learn from the noise and resistant the overfitting case. In the previous works reviewed, the algorithms before fine-tuning have a serious overfitting case the authors faced and used the fine-tuning method to prevent it with a grid search in cross-validation. Decision Tree and Random Forest are two classification algorithms the authors used to random search to apply the classify with a random selection from the training set.

### **2.3 Summary of Authors' Previous Works of Classification in Project Management**

In these previous works researched and reviewed, the authors developed a machine learning model to solve the situation in the project management. The scope and objectives of the authors previous work are archived, and they are beneficial for project planning in project management. In the reviewed, the authors have been preprocessing to prevent the overfitting and noise. After that, they selected a suitable classification algorithm to archive the objectives in higher accuracy. In these previous works reviewed, most of the project management wanted to profitable by used a machine learning from the project. A machine learning is suitable for business project planning to ensure the project profitable and controlled well.

In these learning from previous works, this project must develop more classification algorithms to ensure the outcome is higher accuracy. From the case study, the previous work only used one or two algorithms model to do the analysis for the project management. Expect previous work [10] was used three algorithms to prevent noises and develop higher accuracy outcome. The more algorithms used, the model is not overfitting and the assemble learning method enable select the higher accuracy of the algorithm into the outcome. Additionally on [3] and [6] reviewed summary, the fine-tuning method is required in this research project to prevent the overfitting or underfitting case happened.

As the summary of previous works [5] - [10] reviewed, understand the project objectives is a target to solve the problem of the project management. To success archive the goal, this project must have a clarity data visualization to ensure the dataset is understand and selected the variables to process development a machine learning. The preprocessing must reduce the noises and avoid the overfitting during modelling and postprocessing. In modelling, the model must selected a suitable algorithm to provides a machine learning from the dataset and ensure it is the high accuracy outcome to prove the postprocessing.

In the end, the research project has select a higher accurate algorithm to develop a Machine Learning model and achieve the objective. To this research report, the classification algorithms will develop to achieve the research's objective which are Logistic Regression, Decision Tree, Random Forest and *SVM*.

# Chapter 3

## Research Model

This chapter is introduce a research procedure used develop a Machine Learning classification model in this project. The research model is a methodology used to process through the end of the research project. Each stage of the research procedure is introduced and summarized each step process during the stage.

### 3.1 Research Methodology

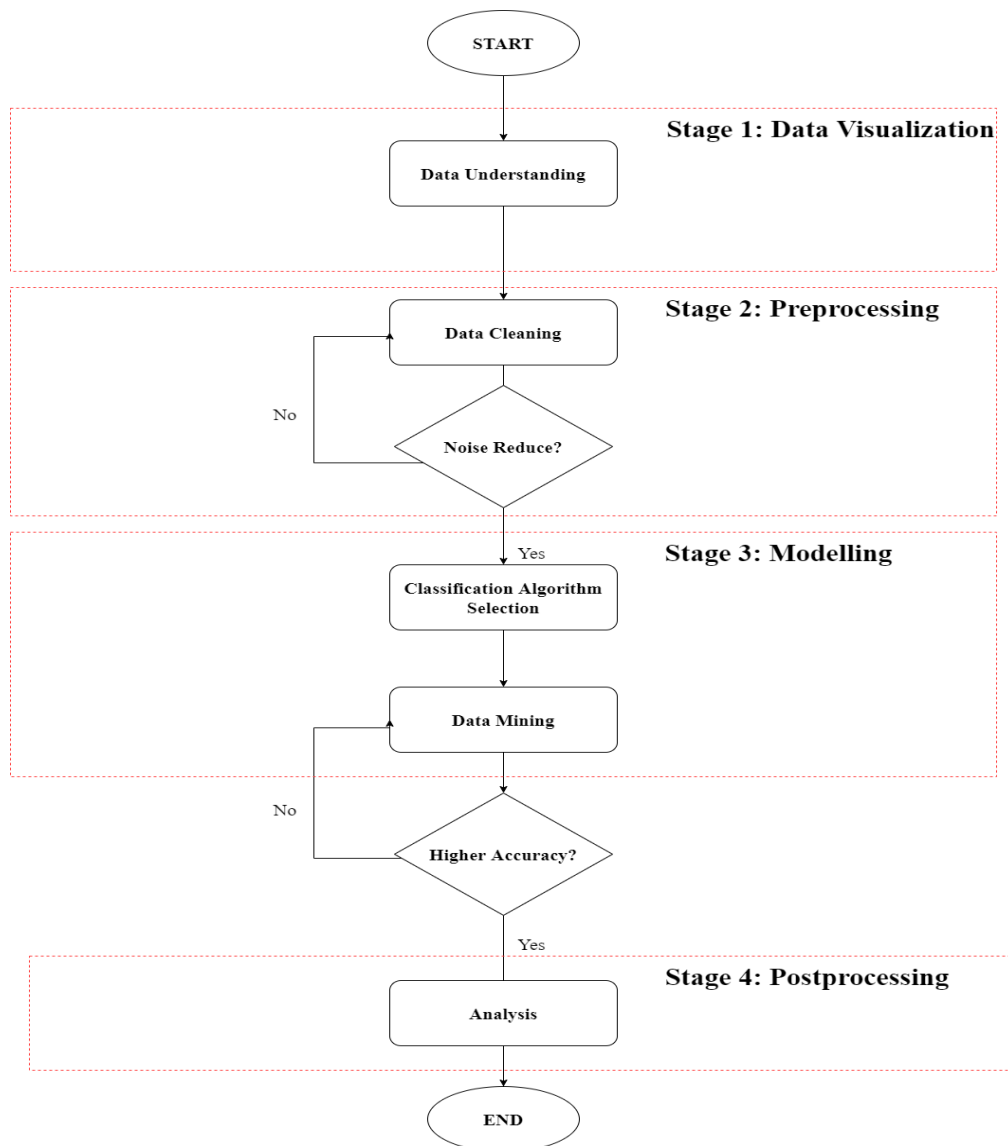


Figure 3.1 Research Procedure

In the figure 3.1, it showed the stage of the data analysis for this project was produced using the research procedure. The research procedure is introduce the stage of the research process to complete the task of this project as below (Section 3.1 – 3.4).

### **3.1.1 Data Visualization**

This stage is understand the dataset of the description and concept. The dataset used was made up of thousands of recorded constructions projects which has included new start, in-progress, and finished. It describes the dataset to understand the meaning each of the data feature to provide the target of project objective. Forwards, it must create a visualization of scatter, bar, and line charts. This data visualization is understand the relationship between two of data features. After the data visualization, the important data features selected to do the feature of the analysis. The sub-stages are data understanding and data feature selection, which are introduce in these below.

#### **3.1.1.1 Data Understanding**

This is a first sub-stage is understand each data feature description from a company's construction project recorded dataset. To understand a dataset, each data feature has list out the description and data type in the research project. Furthermore, the data features will be selected to perform the Exploratory Data Analysis (EDA) process the data understanding by using visualization and object queries, such as scatter, bar, and line charts. The data feature will transform into a visualization to understand the project the company received in next chapter of sub-topic (Section 4.1).

### **3.1.2 Preprocessing**

After data features selected, this stage is a data preparation to clean the data before further develop a classification algorithm. The preprocessing is ensure the data features modify to the machine enable to read and learn on the concept of the data features. The fact of this stage is a machine does not like a human, it unable to read the code character, and a null value in a data feature. This stage is deal with a dirty data and noise, deliver a clear data to a machine. This stage is avoid a machine learning outcome is overfitting and learn from noises. Therefore, the cleaned of dataset will be separated into random split the data to process data modelling training set and test set.

#### **3.1.2.1 Data Cleaning**

Data cleaning is a process to deal a noise from the dataset and clean to the machine. The data quality problem is affect machine enable to read and learn, which are noises, dirty data objects, and duplicate data. Before modifying the data quality, it must identify and understand the dataset quality issue must be faced and select the data cleaning methods to deal the problems. Reasoning the noises data objects can affect a machine cannot learn from the pattern and cause the accuracy outcome in overfitting or underfitting. The data cleaning method used in this research project are imputation, feature creation, label encoder, feature selection and feature scaling. In this summary, this data quality problem must solved and ensure the machine enable to read and learn in the modelling phases to avoid the machine learning problem during modelling phases (Section 4.2.2).

### **3.1.3 Modelling**

This stage is develop the classification algorithms to archive the project objectives. This is ensure it matches up for the meaningful conclusion from the model to dealing with the research objectives. After development each of the algorithms, the fine-tune model is a process to takes a model has trained to tunes or tweaks the model to make it perform a second similar task. The fine-tune model is ensure a model has tuned in best performance to perform a train set outcome.

#### **3.1.3.1 Classification Algorithms Selection**

Before start data mining, the classification algorithms selection is an important sub-stage to select classification algorithms to archive the project objectives. The classification is an algorithm to classify data objects into a multiple class. There are many classification algorithms, and each has its own positive and negative aspects. In the case, the algorithms must be mindfully selected as it influences the model's ability to achieve the project's objective.

#### **3.1.3.2 Data Mining**

After selected the classification algorithms, this sub-stage is used the algorithms selected and to develop a rule of algorithms to enable a machine to learn from the data training. Before process the data mining, the cleaned dataset will be split into training set and test set. Afterward, a machine enables learn the algorithm rule to self-training and analyze the outcome. At the same time, a machine model will be fine-tune with cross-validation and search the great cross-validation to ensure the machine enable analyze more accuracy than the previous outcome. In addition, the fine-tune is to prevent the overfitting or underfitting case.

### **3.1.4 Postprocessing**

The final stage of the research procedure is measure the accuracy between train set and test set. This is ensure a model analysis the actual outcome is higher accuracy than the training set. The overfitting and underfitting is happened when a machine learned from noises and less data. This is a stage to avoid the overfitting and underfitting case. Even it is happened, the stage must return back to modelling phase to deal the case. Therefore, the higher accurate of the model was selected and develop to a construction project management. The higher accurate of the model will analyze to provide used in the project planning to earn profitable from the construction project.

#### **3.1.4.1 Analysis**

After developed a machine model, this stage is used the test set to test the accuracy of algorithms with the training set. This is ensure the model actual predicted the accuracy higher than the training set. This is to avoid the model in underfitting or overfitting case happened. This step is analysis the accuracy of algorithms with confusion matrix and select the higher accuracy of the algorithm model develop to the construction project management.



### **3.2 Summary on Research Model**

This chapter is summarized the process of the data analysis by using research methodology. This research methodology can be explain a machine enable learning the dataset concept and data mining with the algorithm rules. Besides, the most important phase in the research methodology is preprocessing. The reasoning behind of the phases is the dataset must reduce the dirty data to prevent the machine learning from the noises and resulting the accuracy of the outcome is affected in overfitting. It is likely that a machine cannot learn from the human work or idea, the dataset must have generate to ensure the machine is enable to read and learn with the algorithm rules.

The data understanding of the construction project recorded has been understood. As knowledge of the understanding, these data features descriptions (Section 3.3) are represented each project of the information. In construction project recorded summary, a project has the own project type and stated the information of project received. In the same case, each of the project has different status. Each of the project has the DSF numbers which the construction project registered the safety framework from the government. Besides that, some of them remain merely new-start, some are in-progress, and some have already been completed. Additionally, some of the projects new-start and in-progress don't have the actual cost spending recorded. The reasoning is the projects is ongoing and the project manager enable to estimate the cost spending. In the project completed, the project manager will recorded the actual cost spending and project duration in the dataset. According to the research project objectives, this research must look forward on the completed projects into data analysis as achievable on the goal.

Following the data features understanding on sub-topic (Section 3.3), this research project should looking on the completed project records. The completed project has record the project actual end date and the total phase actual spending cost amount. It can know the actual end date is not due on the time in planned end date and result the cost will be spend on time extended. The furthermore detail will be introduced in Chapter 4.

# Chapter 4

## Data Understanding

In this chapter, the dataset has been understood and a meaningful data feature is taken into a graph visualization to understand the frequency of the different types of construction projects which the company has taken. Hence, the summary of the data understanding and data visualization is introduced in this chapter.

### 4.1 Introduction of Dataset

The dataset chosen for this project is a construction project planning dataset which it is recorded projects the company received. The dataset was taken from the open-source dataset internet, and it allows to be used for the research project purpose. In the dataset, it has 11793 projects recorded and each project has 14 data features: project geographic district, project building identifier, project school name, project type, project description, project phases, project status, project phases actual start date and planned end date, project actual end date, budget amount, final estimate of actual costs, the total phases actual spending amount, and DSF numbers.

In the dataset quality, there are some of the data features that do not have a value. It can present some of the projects that do not have planning complete or it is ongoing progress. Besides that, some ongoing and starting up projects are not filled on the estimate of actual costs in the end of phase. In summary, the class is unlabeled to let the machine learning and. In the case, it must create it into a label class to let a machine be able to learn the dataset. This dataset quality issue will be introduced in the next chapter of data preparation (Section 5.1.1).

## 4.2 Data Features Description

In the dataset, the project record has contain 14 data features to present the project planning information. The data features are an information to present the meaning of the project record detail. The data features help to understand the project planning information and it presents the project on the actual spending budget. Therefore, the data features are introduction as listed in this below.

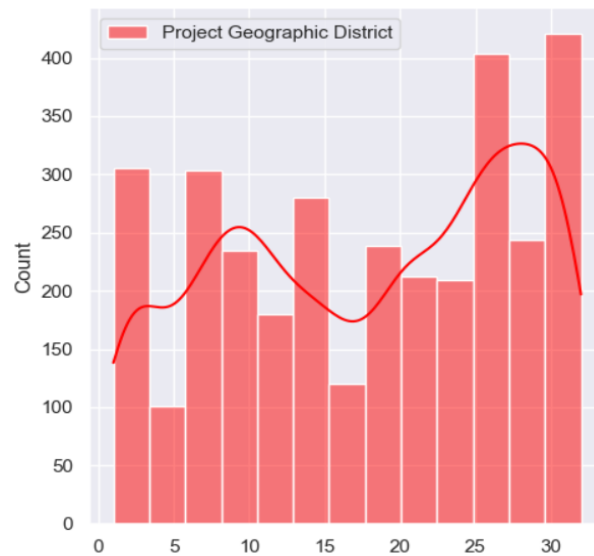
- **Project Geographic District (Ordinal).** An identifies location geographic of the project. It consists of a number to present the meaning of the location geographic of the project (e.g., 1 – New York City, 2 - California).
- **Project Building Identifier (Nominal).** An identifies of the project code. It presents the code of the project. The project code combines a letter with the three numbers (e.g., A069).
- **Project School Name (Nominal).** The project name. There are many projects present the name to understand the title of the project planning.
- **Project Type (Nominal).** It represents type of the project. There are many different types of the project. The project type of code consists of the letters (e.g., SCA CIP).
- **Project Description (Nominal).** It presents the project work detail. It is present the further information from the project school name. The project school name is a title of the project.
- **Project Phase Name (Nominal).** It presents the phases of the project. Each of the project has been recorded the different phases. The phases of the project are construction, purch and install, scope, design, construction management (CM) and fixture.

- **Project Status Name (Nominal).** It presents the status of the project. Each of the project phase has the status to understand the progress of the planning. The status name of the projects recorded are project new start (PNS), in-progress and complete.
- **Project Phase Actual Start Date (Nominal).** It presents the actual project start date. The data information is consists of the format of the start date is month, date, and year (e.g., MM/DD/YYYY, 12/31/2023)
- **Project Phase Planned End Date (Nominal).** It presents the project work planned the project end date. It consists of the format of the planned end date is month, date, and year (e.g., MM/DD/YYYY, 12/31/2023). Some of the information are put the codes letters, which are DIIR, DOES, FTK, DOER, IEH, DOEL and TPL. These codes are present the project manager doesn't put the amount values and replace with the codes.
- **Project Phases Actual End Date (Nominal).** It presents the project work actual end date. The data information is the project actual work end finish of the date. It consists of the format of the end date is month, date, and year (e.g., MM/DD/YYYY, 12/31/2023)
- **Project Budget Amount (Ratio).** It presents the client spend the total budget cost on the construction project. Some of the information are put the codes letters, which are DIIR, DOES, FTK, DOER, IEH, DOEL and TPL. These codes are present the project manager doesn't put the amount values and replace with the codes.
- **Final Estimate of Actual Costs Through End of Phases Amount (Ratio).**  
It presents the project budget estimate the costs spending through the end of phases. The costs are the project management to estimate the actual costs until the end of the phases. It consists of the number of total estimates of the budget amount.

- **Total Phase Actual Spending Amount (Ratio).** It presents the actual budget on a phase in actual costs. The costs are the project is fully complete and the actual costs spending is made. It consists of the number of actual budget spending amount
- **DSF numbers (Nominal).** It presents the Dam Safety Framework (DSF). The construction project has the own DSF code to present the safety framework in the project planning. The project of the DSF is consists of the code which is DSF letters with random number (e.g., DSF000123456789). Each of the project has contain many different types of DSF numbers registered.

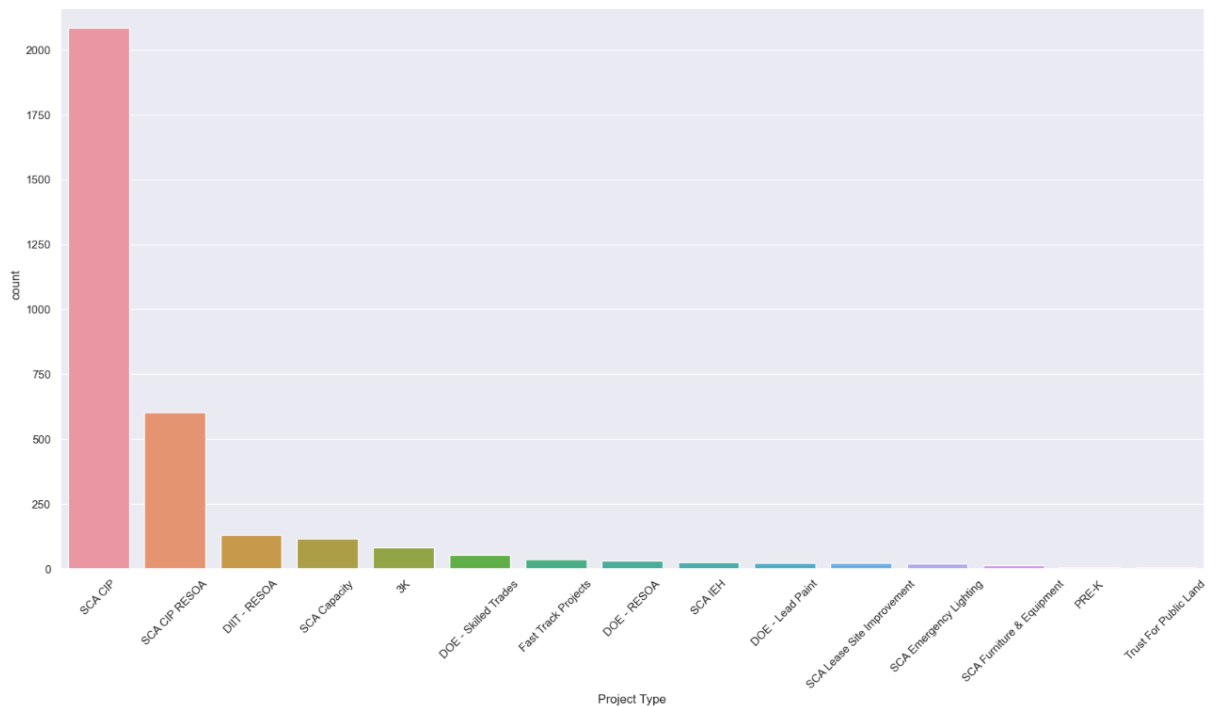
### 4.3 Data Visualization

This topic is introduce the data visualization of the construction project records dataset. The data visualization is a visual tool to represent the data in graphics, which are charts, plots, bar, and infographics. The visual displays the information of the data feature relationship and data feature information. In the data visualization, the projects are visualized in 3253 completed projects. The completed projects have present the full detail of the company received.



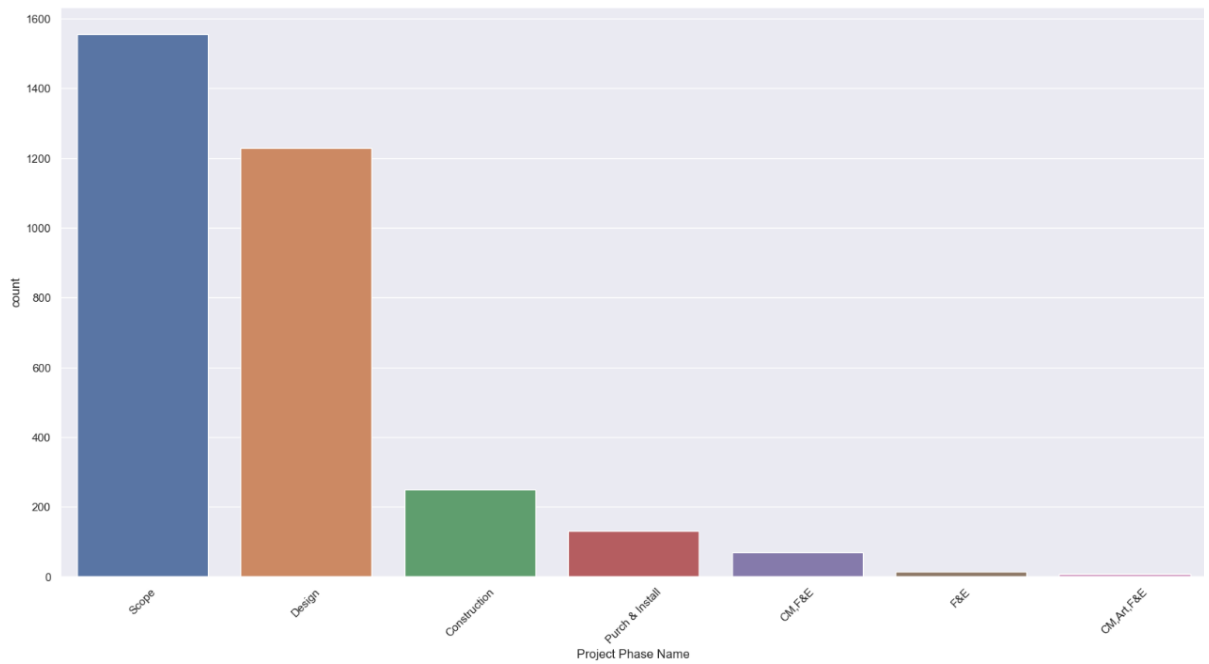
*Figure 4.1 Project Geographic District in project records*

In figure 4.1 shows the completed project of the geographic district the company taken. In continuous bar graphs, the company most received geographic district of the project within code 30 and 35. Furthermore, there are two peaks in the trend line, which are geographic district code 10 and 30. It mean most of the projects the company received frequency are code 10 and 30.



*Figure 4.2 Project Type of project records*

The construction company has received 15 types of projects. In these 15 projects type the company received, the company most received the project type is code SCA CIP which is 2087 projects. The code SCA CIP is the first place the company most likely received this project type. In addition, the project type codes SCA CIP RESOA and DIIT-RESOA, which are 602 and 131 respectively, are in second and third place. In this bar chart visualization, these are the three projects type the company mostly received from the clients.



*Figure 4.3 Project Phase Name*

The summary of figure 4.3, the construction company received different phases of the project from the client's request. In the figure summary, the company most received 1555 projects is scope project. These construction projects are to scope the project planning and design the concept of the construction place before start construct the building. Aside from that, the design is a second most project phase the company received which is 1228. The design is the project manager design the building construction and analysis the materials used in the construction planning in planning management. The third most project of the project phase the company received is the construction. The construction phase is to construct the project that will be built into the building.



#### **4.4 Summary of Data Understanding**

The dataset is collected from the internet source. Those data features are not introduced from the internet source. To solve this problem, the dataset must overall and analysis before starting the data understanding stage. These are data features summarized are meaningful to understand the record of the construction project frequency received. Additionally, the original dataset is recorded 11794 projects and those records are recorded in new start, in-progress and complete. To achieve the research objectives, the records must looking on the completed projects the company taken. The completed project has completed recorded in each of the data features. In the end, the total of the completed project is analyzed for this research project is 3253.

From data visualization (Section 4.3), it can notice that there are only three data features are selected make into the data visualization. These are three data features is a perfect attribute set from the original dataset. Furthermore, there are some of the data features contained dirty data or unnecessary learn. It is unable let a machine to identify learning from the dataset and impossible to deal it into a data visualization. This data understanding stage only introduced the dataset collected and cannot make a data cleaning in this stage. Hence, the data visualization (Section 4.3) is limited and must take the completed data feature from the data description (Section 4.2). In the conclusion of this summary, the data cleaning must deal with the data cleaning methods in the next chapter.

# Chapter 5

## Data Preparation

This chapter introduce the identify the dataset quality problem and faced must been cleaning with using data cleaning methods. Afterward, the cleaned dataset will slip into training set to ensure machine learning used algorithm rules to scale and learn. The preprocessing progress are summary in this chapter.

### 5.1 Data Preprocessing

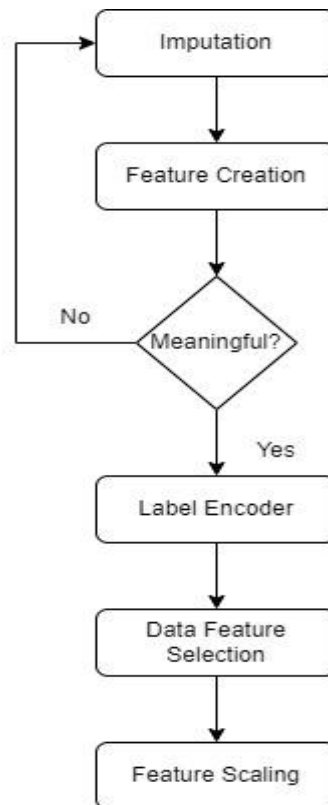
Preprocessing is a phase of the research methodology to transformation the data to ensure a machine enable to training the cleaned data with preprocessing technique. This phase is important to affect the training model and test model. To clean the data, the preprocessing must be faithful to the research objective. This topic introduces the dataset of problem quality will be encountered in the research project (Section 4.2.1) and the technique used to data cleaning in preprocessing (Section 4.2.2).

#### 5.1.1 Data Quality Problem

From the data understanding, there are a lot of dirty data can be affect the accurate of the machine learning. According to the data features description (Section 3.3), there are two data features have a code to replace the unknown amount to record the project budget, which are project budget amount and final estimate of actual costs through end of phases amount. The codes are not represent detail from the internet source and assume it is an unknown budget amount. Aside, there are three data features are recorded in date format, which are project phase of actual start, planned end and actual end. The project manager records the data format in the completed project.

Besides that, the dataset of the data feature does not contribute significantly to the success of the company's overall project. It can create a new data feature to present more meaningful information from two related data feature on the dataset. Furthermore, there are many data features are affect a machine to unnecessary to learn with the algorithms rule. The machine learning hard to learning from the dataset with algorithm.

### 5.1.2 Data Cleaning



*Figure 5.1 Flowchart of the data cleaning*

The figure 4.4 above shows the flowchart of the data cleaning methods used in the data cleaning phase. The dataset has serious data quality issue and must deal with the manual data cleaning methods. These are data cleaning approach is followed and process longer times to ensure the machine has a good algorithm learning with cleaned data. These are the data cleaning methods used in this data cleaning phases are imputation, feature creation, label encoder, data feature selection and feature scaling.

Imputation is a method used to replace a dirty data or incomplete data. In this dataset, there are two data features are Project Phase Planned End Date and Project Budget Amount have content few dirty data. These recorded code 'DIIR', 'DOES', 'FTK', 'DOER', 'IEH', 'DOEL', and 'TPL' are not present meaningful information in the dataset. In Project Budget Amount, the imputation can replace these dirty data into the meaningful data with a mean. Reasoning to understanding the average of client spending and the preservation of summary statistics in relation to project budget is the justification for choosing a mean to replace the dirty data.

It is successful to imputation the dirty data in Project Budget Amount. However, the Project Phase Planned End Date does not take the place of the recorded code. To deal with this, the feature creation method is used to create new data feature of the actual and planned duration. The dirty data must set into the null object data to prevent the new data feature doesn't content the dirty data. The Actual Duration is a new data feature created content an actual construct day duration from actual start date and actual end date. In the same case, the Planned Duration is a new data feature created content a planned construct day duration from actual start date and actual end date from actual start date and planned end date. Deal with the null object data of Planned Duration, the step has loop back to the imputation method to replace the null object data with the mean. Reasoning is it presents an average the project manager mostly to plan to finish the project within the planned. After that, the dataset is not present more meaningful information to classification with Machine Learning.

After verifying the dataset's information, feature creation is processed to create a new data feature to present a meaningful of the dataset. The new data feature that was produced as the outcome of a relationship between two data features. In the research project, the Duration Left, Overweight Duration, Remain Amount, Acceptance Project and Overweight Spend are created in the dataset. In summary of new data feature created, the Duration Left is the ratio data which present a remaining time for a project that was completed between the planned and actual end dates. In the case, the Overweight Duration is a nominal data created to present a project due overtime from the Duration Left. The Overweight Duration is present positive if the negative value is content in the Duration Left. Besides, the Remain Amount is a ratio data created to present the company earn the profit or loss from the project budget amount. Forwards, the Acceptance Project is a nominal data created to present the acceptance project the company can receive from the profit or loss project. Lastly, the Overweight Spend is a nominal data created to present a project overweight spend in a project from the planned and actual spend budget amount.

After feature creation, the label encoder is a data cleaning method to transform a nominal data into an ordinal data. The nominal data is present an object data with the strings, and the ordinal data is present a ranking of object data. In the dataset, there are

data features set in the nominal data and it must be transform into an ordinal data to ensure the Machine Learning enable to learning with algorithms. In this research project, the label encoder must process with manual, and reasoning is the dataset has recorded many of project type and project phases. Plus, the Overweight Duration, Overweight Spend and Acceptance Project are a data features which present true or false. These are the data features must be label encoder to ensure the Machine Learning enable to readable the data.

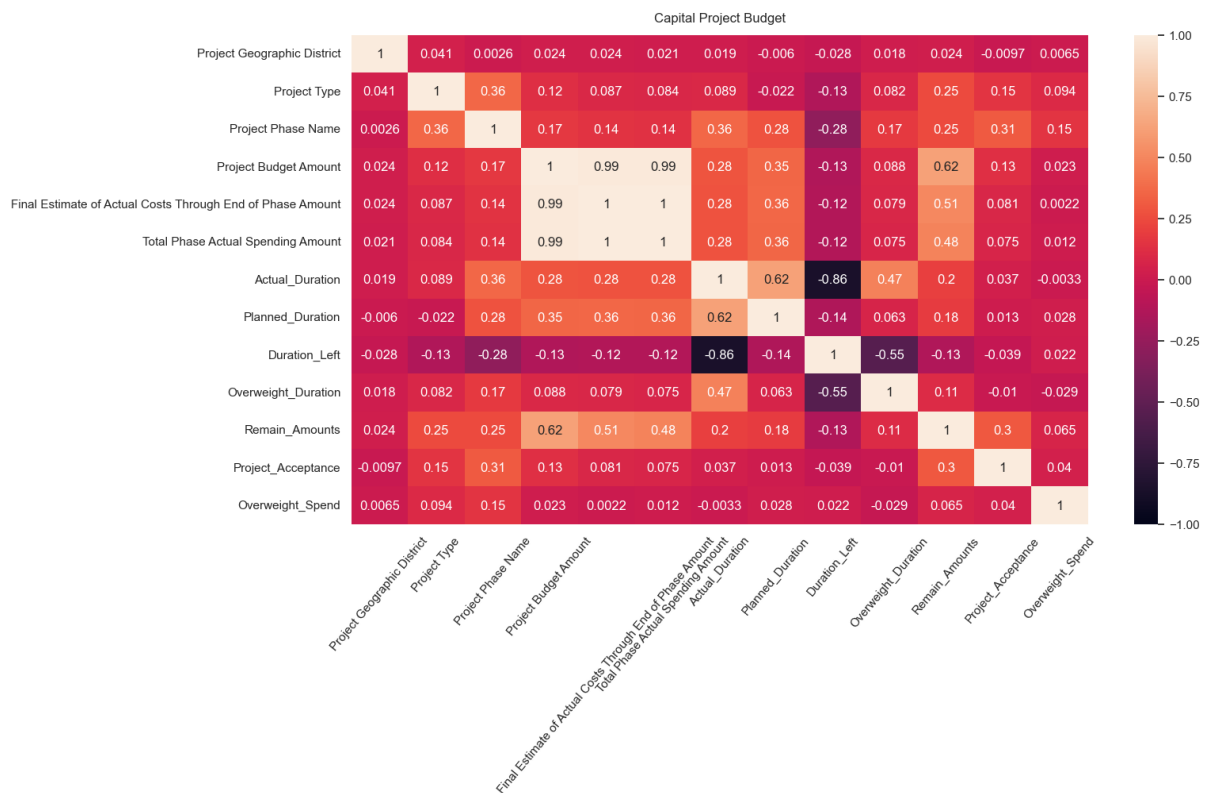


Figure 5.2 Feature Selection with Correlation Matrix

Process on the feature selection, this method is to drop the unnecessary data feature to avoid the machine to learning the noise of the data feature and affect the algorithms accuracy of the outcome. To deal with this, the correlation matrix is a way to selection the meaningful of the data feature for data classification. The correlation matrix is represent the relationship between each of the data features' correlation coefficients. In the figure 4.5 shown the correlation matrix visualization created in this research project. In summary, the weakest relationship is Project Type and Project Phase Name which is 0.36. It means these both data feature is small affect increase in 0.36 correlation value.

In the feature selection must select the relationship data feature which is higher than 0.5 correlation value with correlation matrix. As a result, Figure 4.6 below shows the correlation matrix's final result after feature selection.



Figure 5.3 Correlation Matrix after Feature Selection

In analysis, the Actual Duration and Planned Duration has a great relationship in 0.62 correlation value. Additionally, the Remain Amount and Project Budget Amount has a great relationship in 0.62 correlation value. Besides, these are the stronger relationship are the Project Budget Amount, Final Estimate of Actual Costs Through the End of Phase Amount, and Total Phase Actual Spending Amount, which is between correlation value in 0.99 and 1. It is mean whatever a data feature is increase, rest of the data features are serious affected increase. As the classification learning, the remain amounts is taken as the classification learning to the machine. To enable to machine readable, the remain amount of the data object has changed into ordinal as true or false. The data object's remaining amount is true when the value is positive and false when the value is negative.

Feature scaling is a data cleaning method to adjusting the value of different ingredients in a recipe to ensure the data feature is balanced. It helps the data features with different scales measure together with classification algorithms, and prevent the overpowering and affect the accurate outcome. After feature selection, the duration and budget amount are different measure units. Duration is recorded in days unit and the budget amount is recorded money unit. Without feature scaling, the variations in feature magnitudes can cause machine learning algorithms to converge incorrectly from noises.

In the end, the research project is selected the Remain Amount to the machine to learning the determine of the project profitable as the research project's objective. To ensure the data feature is enable the machine the ability to learn, the data feature was transaction into a label data object. In summary, the positive values in the Remain Amount means the project have remain budget from budget amount and actual spending and it is profitable. The positive value has change into "Yes". In the same case, the negative value means project have not profit and loss for company. The negative value has change into "No".

## **5.2 Split Data Training and Test Model**

These are the data cleaning methods used to increase the dataset quality in the preprocessing phases (Section 4.2). After preprocessing, the dataset must be split into the training set and test set. The training set is let the machine to learning and training with the algorithms rules. Furthermore, the test set is used to test the accurate of algorithms to ensure the Machine Learning model is not overfitting or underfitting case happened. The overfitting or underfitting case happened in test case, resulting the algorithms model is necessary return to training model to do hyper tuning the parameters of the algorithm to prevent the overfitting or underfitting case. In the research project, the dataset is split into 60% of training set and 40% of test set Reasoning is ensure the Machine Learning model focuses more on the testing model and less on training model with the algorithm rules. Resulting is the Machine Learning present the actual outcome to the project manager. They have a better analysis from machine outcome and making decision in the project planning management.



# Chapter 6

## Algorithm Selection and Performance

### Evaluation

This chapter is summary a classification algorithm selected to developed in the modeling stage. Furthermore, each of the algorithm will introduced the regular performance and the fine-tune of the search the best parameter. The postprocessing introduced and analyzed the higher performance of algorithms used develop to Machine Learning Model.

#### 6.1 Classification Algorithm Selection

As towards the research objectives, the classification algorithms developed in this project include Logistic Regression, Decision Tree, Random Forest and SVM. The machine must analyze the project's budget to determine the received projects are profitable for the construction company. The reasoning these algorithms were created for this research project due to the dataset contains label data and it is a suitable to select the supervised learning algorithms to developed (Section 5.1.2).

#### 6.2 Modelling

In the modeling stage, the cross-validation is developed in these algorithms to prevent the overfitting case happened in this research project. The cross-validation is setting on 5 to ensure the machine is learn the medium-size training set and maintain the performance learning practice of outcome with balance bias and variance. Therefore, it can be easier to fine-tuning the parameters of each algorithm in the learning. After developing the algorithm, the fine-tuning parameters is necessary to tuning the algorithm to increase the performance accuracy and prevent overfitting case which compare with the test model.

##### 6.2.1 Logistic Regression

In this algorithm, the learning is set to 1000 as the maximum of iterations with a random state of the sampling in 42. The resulting the accuracy of this algorithm with the cross-validation is 71% and this is a good performance algorithm in the training model.

Unfortunately, the algorithm must fine-tuning the parameters to increase the accuracy of this algorithm and reduce the higher accurate of the learning. Resulting in the grid search in the parameters with cross-validation, the best parameter in this algorithm is  $C = 3$ , penalty = 'L2' and solver = 'lbfgs' and the accuracy performance is 85%

### **6.2.2 Decision Tree**

In the training, the accuracy of the algorithm in random sampling in 42 of the Decision Tree is 94%. It is shown this algorithm has a good performance in this training model. However, the test model of this algorithm learning is overfitting. The accuracy of training learn is higher than the actual learn. It is mean the algorithm is unstable to learn in actual test model. Therefore, it is necessary to the fine-tuning the parameter to reduce the learn and prevent the overfitting case.

After the grid search with cross-validation, the best tuned parameters of the algorithms search are criterion = 'gini', maximum of depth = 5, maximum of features = 'sqrt', minimum of samples leaf = 10, minimum of sample split = 5, splitter = 'best' and result the accuracy is 80% with this algorithm.

### **6.2.3 Random Forest**

This algorithm is the machine use a random sampling 42 taken from Decision Tree to classify the outcome. In the training model, the accuracy performance of the algorithm is higher than another algorithm which is 94%. Even fine-tuning the parameters in this algorithm, the accuracy performance is same with the test model. To handle this situation, it must grid search the best parameters to prevent the more learning with the algorithm rule. The grid search is used similar from the Decision Tree (Section 6.2.2) with 20 trees. After the grid search, the result 90% accuracy performance of the best tuned parameters are criterion = 'gini', maximum of depth = 5, maximum of features = 'None', minimum of samples leaf = 10, minimum of sample split = 5 and number of estimators = 20.

#### 6.2.4 SVM

This algorithm model is used a street the measure a maximal margin to classify the label data. The result of the accuracy develop with this algorithm is 66% and it is a lower accuracy performance with another algorithm. Additionally, it shows the accuracy performance of test model is lower than the training model. It is overfitting serious happened and provide use fine-tuning to grid search the best parameters in this algorithm. To increase the performance, the best parameters tuned are  $C = 2$ ,  $\gamma = 0.1$ , kernel = 'linear'. Resulting it is suitable used the hard margin to classify the outcome with this tuned parameter and the accuracy performance is increase into 80%.

### 6.3 Test Set Model

The 40% of the dataset was taken into a test set to test an accuracy of the algorithms in actual learning. In this research project, the overfitting case is happened in the test model and necessary to solve it with fine-tuning parameter of algorithm (Section 6.2). As shown in table 6.1 below, Model 1 represents the accuracy of Machine Learning using algorithm rules in the training model, Model 2 represents the accuracy following grid search with the best parameters tuned, and Model 3 represents the accuracy in the test model.

| <b>Algorithms</b>          | <b>Model 1</b> | <b>Model 2</b> | <b>Model 3</b> |
|----------------------------|----------------|----------------|----------------|
| <b>Logistic Regression</b> | 71%            | 85%            | 86%            |
| <b>Decision Tree</b>       | 94%            | 80%            | 81%            |
| <b>Random Forest</b>       | 94%            | 90%            | 90%            |
| <b>SVM</b>                 | 66%            | 80%            | 85%            |

*Table 6.1 Accuracy Performance of Each Algorithms*

This research project must tuned the parameters of the algorithms to ensure it must develop a best performance accuracy of Machine Learning in training set and test set. Summary of table 6.1, the Random Forest is the highest accuracy in three model. Even it tuned best parameters from grid search, the Model 2 and Model 3 is same accuracy in fine-tuned training and test performance. Besides that, *SVM* is a weakest performance in Model 1. However, it shows the performance in learning is best that *SVM* is a best learning in Model 2 and Model 3 with hard margin classify. In the end, the arrangement accuracy performance of the Machine Learning model with algorithm in this research project is Random Forest > Logistic Regression > *SVM* > Decision Tree.

As summary on the accuracy performance, the behind of the accuracy performance of the algorithm is measure with the Confusion Matrix. Therefore, the Confusion Matrix is necessary to develop and summary the performance behind the accuracy in this research project. Hence, each algorithm of the result in training and test is introduced and summary the performance with Confusion Matrix (Section 6.3.1-6.3.4).

### 6.3.1 Logistic Regression of Test Performance

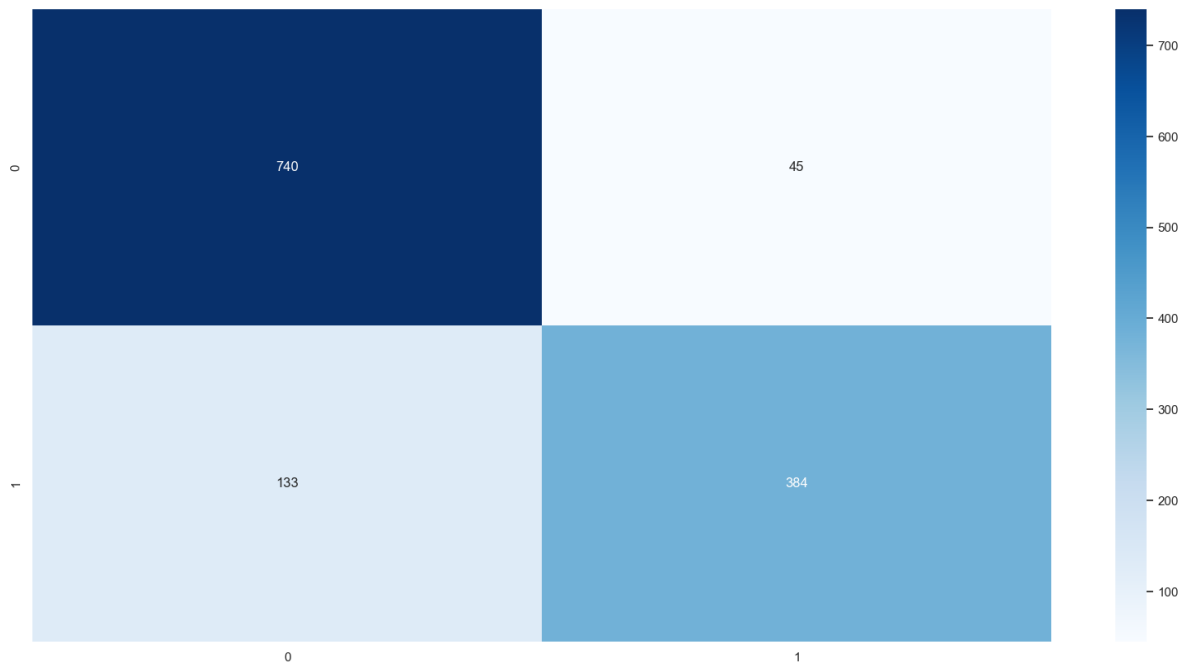


Figure 6.1 Confusion Matrix of Logistic Regression

| Class                   | Precision | Recall | F1-score | Support |
|-------------------------|-----------|--------|----------|---------|
| <b>0</b>                | 0.85      | 0.94   | 0.89     | 785     |
| <b>1</b>                | 0.90      | 0.74   | 0.81     | 517     |
| <b>Accuracy</b>         |           |        | 0.86     | 1302    |
| <b>Macro Average</b>    | 0.87      | 0.84   | 0.85     | 1302    |
| <b>Weighted Average</b> | 0.87      | 0.86   | 0.86     | 1302    |

Table 6.2 Score Matrix of Logistic Regression

In the Confusion Matrix of the Logistic Regression, the accuracy of the test performance is 0.86. Both classes show there are good precision in the 1302 test set size and result this algorithm is precision the actual outcome. However, the negative class has stronger recall performance than the positive class. As the performance of the Logistic Regression, the precision and recall are balance which the average of F1-score is 0.85.

### 6.3.2 Decision Tree of Test Performance

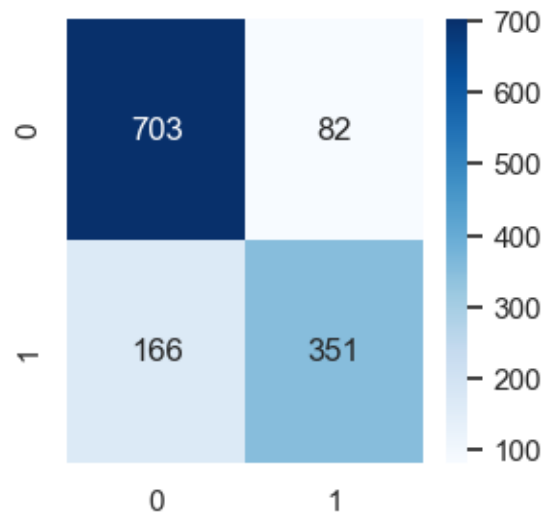


Figure 6.2 Confusion Matrix of Decision Tree

|                         | <b>Precision</b> | <b>Recall</b> | <b>F1-score</b> | <b>Support</b> |
|-------------------------|------------------|---------------|-----------------|----------------|
| <b>0</b>                | 0.81             | 0.90          | 0.85            | 785            |
| <b>1</b>                | 0.81             | 0.68          | 0.74            | 517            |
| <b>Accuracy</b>         |                  |               | 0.81            | 1302           |
| <b>Macro Average</b>    | 0.81             | 0.79          | 0.79            | 1302           |
| <b>Weighted Average</b> | 0.81             | 0.81          | 0.81            | 1302           |

Table 6.3 Score Matrix of Decision Tree

In the Confusion Matrix of the Decision Tree, the accuracy of the test performance is 0.81. In the result of score matrix, both classes have a good precision, and the average precision is 81%. In the summary of both classes, the negative class of the recall score is higher than the positive class. Resulting the negative class is a best performance than positive class which the negative class of the F1-score is higher than the positive class.

### 6.3.3 Random Forest of Test Performance

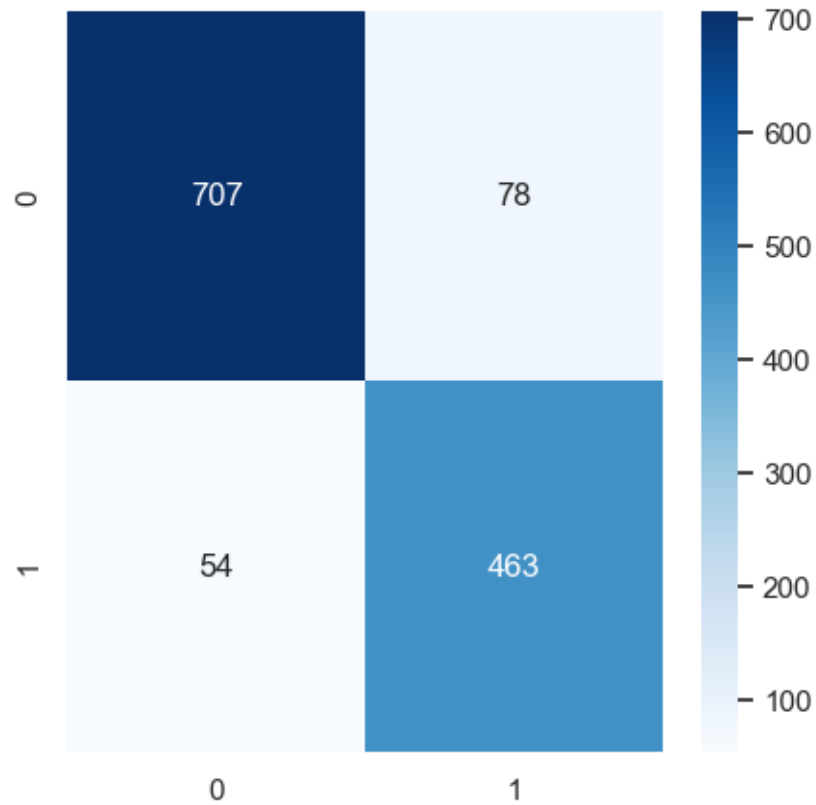


Figure 6.3 Confusion Matrix of Random Forest

|                         | Precision | Recall | F1-score | Support |
|-------------------------|-----------|--------|----------|---------|
| <b>0</b>                | 0.93      | 0.90   | 0.91     | 785     |
| <b>1</b>                | 0.86      | 0.90   | 0.88     | 517     |
| <b>Accuracy</b>         |           |        | 0.90     | 1302    |
| <b>Macro Average</b>    | 0.89      | 0.90   | 0.89     | 1302    |
| <b>Weighted Average</b> | 0.90      | 0.90   | 0.90     | 1302    |

Table 6.4 Score Matrix of Random Forest

Summary of the score matrix of Random Forest, it shows the best performance in the Machine Learning with this algorithm. Reasoning it has best precision, recall, and F1-score for both classes. Therefore, resulting the accuracy of this algorithm is 90% which is higher than another algorithm in this Machine Learning model.

### 6.3.4 SVM of Test Performance

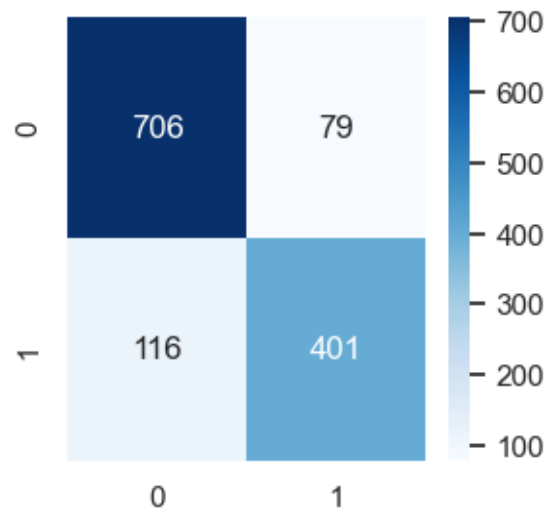


Figure 6.4 Confusion Matrix of Logistic Regression

|                         | <b>Precision</b> | <b>Recall</b> | <b>F1-score</b> | <b>Support</b> |
|-------------------------|------------------|---------------|-----------------|----------------|
| <b>0</b>                | 0.86             | 0.90          | 0.88            | 785            |
| <b>1</b>                | 0.84             | 0.78          | 0.80            | 517            |
| <b>Accuracy</b>         |                  |               | 0.85            | 1302           |
| <b>Macro Average</b>    | 0.85             | 0.84          | 0.84            | 1302           |
| <b>Weighted Average</b> | 0.85             | 0.85          | 0.85            | 1302           |

Table 6.5 Score Matrix of SVM

Summary the score Matrix of SVM, the accuracy performance of this algorithm is 85% that it is a good learning in the test model which compared with training model (Section 6.3). Both classes have a good precision in the actual learning. However, the negative class of the recall score is higher than the positive class. This algorithm has a good F1-score which is 84%. Resulting both classes can classify the outcome in balance.



#### **6.4 Summary of Postprocessing**

In postprocessing, it must selected the higher performance of the algorithm to provide the Machine Learning model to develop for the construction project company. As the training and test model result (Section 6.3), the Random Forest is selected for development the Machine Learning model in the project planning management. Reasoning the Random Forest is higher accuracy performance than other algorithms. In Model 1, the accuracy performance is similar with the Decision Tree. After tuned best parameters with grid search, the accuracy performance is higher than other algorithms in Model 2. In the case, the performance in the training and test model is similar with the Model 2 (Section 6.3).

In the Confusion Matrix of the Random Forest (Section 6.3.3), it has best performance in precision and recall. In the same result, it has good F1-score that it is balance to classify the positive and negative class. In the parameter of Random Forest, it is taken 20 trees to learning the important data features to process the classify the class. As a result, it has the best learning performance and allow to require fewer trees to be learned. In the summary of this research project development, the Random Forest is suitable for Machine Learning model to the construction project company to analysis the remain budget earned from the received project.

# Chapter 7

## Conclusion and Recommendations

Since this is the end of the research project, this chapter is introduced to conclude this work and summarize the idea in the end of research reports. Furthermore, this chapter brings the idea on the recommendation to improve this research project. The ideas can be improved for the future development and case study case expansion.

### 7.1 Conclusion

In this research project, it is success achieved the objectives to develop a Machine Learning model for classification in construction project planning. The construction project manager enable used the Machine Learning model to analysis the project profitable with the parameters inputs. The parameters inputs can be analyzed are duration and cost estimates budget to classify the project profitable. The best performance of the Machine Learning model is reduce the human mistake in project planning and understanding the development duration affect the project budget spending and profit.

The previous works reviewed of this research project, the authors were proof and developed the Machine Learning model to improve the business company productivity and profitability in the business analysis. From the previous works reviewed, [6] and [9] are the good reviewed that the authors achieve the real-world business objectives. The [6] has summarized the fine-tuned parameters by grid search with the cross-validation. It helps this research project understand the reason of the grid search in hyper tuning. Furthermore, the [9] has a good dataset the authors used which the construction project has bring the risks to affect the project budget and bidding price. It means that the construction project must has perfect information to observation before planning the project budget with client.

As designed the research model, it is success planned to develop a Machine Learning model with the hardware and software tools, and research methodology. The planning phase is important to analysis requirements to develop a Machine Learning. Besides, the research methodology is planned to design the process of the phases and follow

through. Resulting the hardware and software tools success help this research project develop a Machine Learning model.

In summary of the dataset used, the dataset is great used to develop for a Machine Learning model. However, the dataset lacks clarity and contains a lot of noises to data cleaning challenging in preprocessing. After correlation matrix, it found out there are a lot of data features are not related relationship and difficult let a machine to learning with algorithms. Resulting the accuracy of the Machine Learning is higher and shown the overfitting case.

After challenge through the research objectives, there are supervised learning algorithms are selected to develop for a machine to learn and test the before selected the higher performance algorithm. In the end, it found out Random Forest is suitable to classify the project budget for Machine Learning model. The accuracy of the Random Forest is definitely higher than other algorithms. Therefore, Random Forest is a higher performance to classify the project budget earned in training and test model.

## 7.2 Recommendations

As a conclusion to the work done on this research project, there are recommendations that can improve the quality of the research project. The dataset has a quality problem which taken from the internet source. The dataset has contains the codes and it represents nothing. The dataset's author does not mention the meaning of the codes. Resulting the codes are assume it as the dirty data objects on this research project. Furthermore, the dataset presents the construction development duration, and budget cost planned and spending. As the [9] is a best reviewed that the authors used the dataset which has contain risk factors that can happened in the construction development. The dataset must contain meaningful details that can affect the project cost budget to easier explore further objectives.

In the literature reviewed, this research project explores more the previous works that the authors developed a Machine Learning model in the construction project. The objective can be learn from the previous work that carried out the benefit to the construction project company and government project. Nowadays, the professional explores and solve the problems by develop a Machine Learning model as a study case for the literature review. Learn the ideas from the literature review and think the idea to solve the situation. The more literature review found, the more learning and experience gained to develop on the study.

This research project is used the Machine Learning model to classification the construction project planning. It can expend more algorithms allow develop to training the dataset. This is a supervised learning to develop a Machine Learning model. As the research project's title mention used the Machine Learning to develop the classification, it can used the Deep Learning model to develop the classification for this research project. The Deep Learning model is enable a machine to learn furthermore on the dataset and result the outcome accurate is higher than another algorithm. The knowledge is not limited to explore more further on the objective and study Artificial Intelligence used in business field.

## References

- [1] Leelavati, Roopa Krishna, and Chandra, “Artificial Intelligence and Machine Learning for Business,” *ProQuest*, Vol. 26, Iss. S6, Aug. 2022. Accessed: July. 28, 2023. [Online]. Available: <https://www.proquest.com/openview/03e87097497bf4977a89a138b0c83466/1?pq-origsite=gscholar&cbl=38744>
- [2] Choon Sen Seah, Yin Xia Loh, Aman Lew Ka Lock Bin Lew Kat Keong, Min Xuan Chin, Gabeirl Chuen Lio, Mei Kay Lee, Li Bin Lim, and Shin Jer Wong, “The Significance of Technology in Digitalising Malaysia Industries,” *Combines*, Vol. 1, No. 1. Feb. 2021. Accessed: July. 28, 2023. [Online]. Available: <https://journal.uib.ac.id/index.php/combines/article/view/4724/1409>
- [3] Sofiat O. Abioye, Lukumon O. Oyedele, Lukman Akanbi, Anuoluwapo Ajayi, Juan Manuel Davila Delgado, Muhammad Bilal, Olugbenga O. Akinade, and Ashraf Ahmed, “Artificial Intelligence in the construction industry: A Review of present status, opportunities and future challenges,” *Journal of Building Engineering*, Vol. 44. Dec. 2021. Accessed: July. 31, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2352710221011578>
- [4] Massimo Regona, Tan Yigicanlar, Bo Xia, and Rita Yi Man Li, “Opportunities and Adoption Challenges of AI in the Construction Industry: A PRISMA Review,” *Journal of Open Innovation: Technology Market, and Complexity*, Vol. 8, Issue 1, Mar. 2022. Accessed: July. 31, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S219985312201054X>

- [5] Theingi Aung, Sui Reng Lina, Arkar Htet, and Amita Bhaumik, “Using Machine Learning to Predict Cost Overruns in Construction Projects,” *Journal of Technology Innovations and Energy*, Jun. 2023. pp. 2957-8809. Accessed: August. 10, 2023. [Online]. Available: <https://www.jescae.com/index.php/jtie/article/view/511/183>
- [6] Prashna Ghimire, Sudan Pokharel, Kyungki Kim, and Philip Barutha, “Machine Learning-Based Prediction Models for Budget Forecast in Capital Construction,” *Researchgate*, June. 2023. Accessed: August. 10, 2023. [Online]. Available: [https://www.researchgate.net/profile/Prashna-Ghimire/publication/372134247\\_MACHINE\\_LEARNING-BASED\\_PREDICTION\\_MODELS\\_FOR\\_BUDGET\\_FORECAST\\_IN\\_CAPITAL\\_CONSTRUCTION/links/64a5daa295bbbe0c6e16c84d/MACHINE-LEARNING-BASED-PREDICTION-MODELS-FOR-BUDGET-FORECAST-IN-CAPITAL-CONSTRUCTION.pdf](https://www.researchgate.net/profile/Prashna-Ghimire/publication/372134247_MACHINE_LEARNING-BASED_PREDICTION_MODELS_FOR_BUDGET_FORECAST_IN_CAPITAL_CONSTRUCTION/links/64a5daa295bbbe0c6e16c84d/MACHINE-LEARNING-BASED-PREDICTION-MODELS-FOR-BUDGET-FORECAST-IN-CAPITAL-CONSTRUCTION.pdf)
- [7] Wuttiping Kusonkhum, Korb Srinavin, Narong Leungbootnak, and Preenithi Aksorn, “Government Construction Project Budget Prediction Using Machine Learning,” *Journal of Advances in Information Technology*, Vol. 13, No. 1, Feb. 2022. Accessed: August. 10, 2023. [Online]. Available: <https://pdfs.semanticscholar.org/45d1/7525c736383799f11b6b1353ce953c2a3d93.pdf>
- [8] Naveen Kunnathuvalappil Hariharan, “Investigating the determinants of successful budgeting with SVM and Binary models,” *International Journal of Management, IT & Engineering*, Vol.11, Jun. 2021, pp. 2249-0558. Accessed: August. 10, 2023. [Online]. Available: <https://ideas.repec.org/p/osf/osfxxx/xf7ak.html>

- [9] Ye Eun Jang, Jeong Wook Son and June-Seong Yi, “Classifying the level of Bid Price Volatility Based on Machine Learning with Parameters from Bid Documents as Risk Factors,” *Sustainability*, Apr. 2021. Accessed: August. 10, 2023. [Online]. Available: <https://stce.huce.edu.vn/index.php/en/article/view/1235>
- [10] Ching-Lung Fan, “Evaluation of Classification for Project Features with Machine Learning Algorithms,” *Symmetry*, Feb. 2022. . Accessed: August. 10, 2023. [Online]. Available: <https://www.mdpi.com/2073-8994/14/2/372>

# FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

|   |                          |
|---|--------------------------|
| <b>Trimester, Year:</b> Trimester 2, Year 4   | <b>Study week no.:</b> 4 |
| <b>Student Name &amp; ID:</b> Lau Zheng Liang   |                          |
| <b>Supervisor:</b> Cik. Zanariah Binti Zainudin   |                          |
| <b>Project Title:</b> Machine Learning for data classification in construction project planning |                          |

## 1. WORK DONE

- Finished Chapter 1.
- Set Objectives and Scope.
- Research new previous works in literature review.

## 2. WORK TO BE DONE

- Summary the literature review
- Correction the report quality.
- Design idea of research methodology.

## 3. PROBLEMS ENCOUNTERED

- Project objective set as a mindset.
- Dataset quality problem.
- Idea to deal data cleaning.
- Previous works review must research within latest 3 years and related with research project.

## 4. SELF EVALUATION OF THE PROGRESS

- Understand the data cleaning method.
- Learn new data mining technique during literature review.



Supervisor's signature



Student's signature



# FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

|   |                          |
|---|--------------------------|
| <b>Trimester, Year:</b> Trimester 2, Year 4   | <b>Study week no.:</b> 5 |
| <b>Student Name &amp; ID:</b> Lau Zheng Liang   |                          |
| <b>Supervisor:</b> Cik. Zanariah Binti Zainudin   |                          |
| <b>Project Title:</b> Machine Learning for data classification in construction project planning |                          |

## 1. WORK DONE

- Finished Chapter 2.
- Design Research Procedure.
- Summary the literature review and get the idea.

## 2. WORK TO BE DONE

- Data Visualization.
- Quality problem review.
- Set idea for data cleaning in coming preprocessing phase.

## 3. PROBLEMS ENCOUNTERED

- The quality after preprocessing.
- Correlation Matrix to select the data features for modeling phase.
- Idea to deal data cleaning.
- Must achieve to research project's objectives.

## 4. SELF EVALUATION OF THE PROGRESS

- Keep updating the progress to supervisor.
- Deal preprocessing with knowledge and must achieve the research objectives.



Supervisor's signature



Student's signature

# FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

|   |                          |
|---|--------------------------|
| <b>Trimester, Year:</b> Trimester 2, Year 4   | <b>Study week no.:</b> 7 |
| <b>Student Name &amp; ID:</b> Lau Zheng Liang   |                          |
| <b>Supervisor:</b> Cik. Zanariah Binti Zainudin   |                          |
| <b>Project Title:</b> Machine Learning for data classification in construction project planning |                          |

## 1. WORK DONE

- Finished Chapter 3.
- Summary more details of Research Procedure.
- Summary Chapter 3.

## 2. WORK TO BE DONE

- Data Visualization.
- Correction the report quality.
- Set idea for data cleaning in coming preprocessing phase.

## 3. PROBLEMS ENCOUNTERED

- The quality after preprocessing.
- Feature creation for more detail on dataset information.
- Correlation Matrix to select the data features for modeling phase.
- Must achieve to research project's objectives.

## 4. SELF EVALUATION OF THE PROGRESS

- Keep updating the progress to supervisor.
- Deal preprocessing with knowledge and must achieve the research objectives.



Supervisor's signature



Student's signature

# FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

|   |                          |
|---|--------------------------|
| <b>Trimester, Year:</b> Trimester 2, Year 4   | <b>Study week no.:</b> 9 |
| <b>Student Name &amp; ID:</b> Lau Zheng Liang   |                          |
| <b>Supervisor:</b> Cik. Zanariah Binti Zainudin   |                          |
| <b>Project Title:</b> Machine Learning for data classification in construction project planning |                          |

## 1. WORK DONE

- Finished Chapter 4.
- Finished data features' description and data visualization.
- Improve the quality of report.

## 2. WORK TO BE DONE

- Data cleaning in manual.
- Correction the report quality.
- Keep updating the preprocessing to supervisor.
- Summary the data preprocessing in the report.

## 3. PROBLEMS ENCOUNTERED

- The quality after preprocessing.
- Correlation Matrix to select the data features for modeling phase.
- Must achieve to research project's objectives.

## 4. SELF EVALUATION OF THE PROGRESS

- Keep updating the progress to supervisor.
- Research the previous works on cross-validation and fine-tuning.



Supervisor's signature



Student's signature

# FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

|   |                           |
|---|---------------------------|
| <b>Trimester, Year:</b> Trimester 2, Year 4   | <b>Study week no.:</b> 10 |
| <b>Student Name &amp; ID:</b> Lau Zheng Liang   |                           |
| <b>Supervisor:</b> Cik. Zanariah Binti Zainudin   |                           |
| <b>Project Title:</b> Machine Learning for data classification in construction project planning |                           |

## 1. WORK DONE

- Finished Chapter 5.
- Finished modeling and summary the outcome accuracy with algorithms used.

## 2. WORK TO BE DONE

- Deal the overfitting case in test model.
- Fine-tuning the best parameter on each algorithm with grid search.
- Summary the fine-tuning reason.

## 3. PROBLEMS ENCOUNTERED

- Dealing the overfitting in manual.
- Use grid search to find the best parameter for fine-tuning.
- Accuracy must tune over 80%.
- Must achieve to research project's objectives.

## 4. SELF EVALUATION OF THE PROGRESS

- Keep updating the progress to supervisor.
- Understand the reason used cross-validation and fine-tuning in this research project.



Supervisor's signature



Student's signature

# FINAL YEAR PROJECT WEEKLY REPORT

(Project II)

|   |                           |
|---|---------------------------|
| <b>Trimester, Year:</b> Trimester 2, Year 4   | <b>Study week no.:</b> 11 |
| <b>Student Name &amp; ID:</b> Lau Zheng Liang   |                           |
| <b>Supervisor:</b> Cik. Zanariah Binti Zainudin   |                           |
| <b>Project Title:</b> Machine Learning for data classification in construction project planning |                           |

## 1. WORK DONE

- Finished Chapter 6.
- Summary the training and test result after fine-tuning model.
- Select the best performance of algorithms to develop a Machine Learning model.
- Summary the research project.

## 2. WORK TO BE DONE

- Improve the research project quality.
- Design the poster.
- Supervisor checking the research project.

## 3. PROBLEMS ENCOUNTERED

- Summary the accuracy outcome with manual fine-tuning.

## 4. SELF EVALUATION OF THE PROGRESS

- Keep updating the progress to supervisor.



Supervisor's signature



Student's signature

# POSTER



## BACHELOR OF INFORMATION SYSTEM (HONS) BUSINESS INFORMATION SYSTEM

### MACHINE LEARNING FOR DATA CLASSIFICATION IN CONSTRUCTION PROJECT PLANNING

AUTHORS  
Lau Zheng Liang 17ACB05620

SUPERVISOR  
CIK Zanariah Binti Zainudin

#### INTRODUCTION

This research project is developed a Machine Learning Model to classify the project budget to the construction project planning. This project achieve the construction project company use a Machine Learning model to profitable and work efficiency in planning management

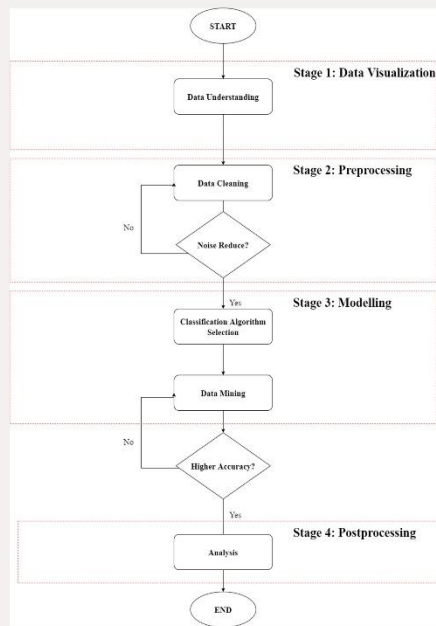
#### OBJECTIVE

- Find a Dataset related with Construction Projects
- To Enhance Data Preprocessing for Improved Analysis and Insights
- To Classify the Project Remain Budget
- Use Fine-Tuning for Optimal Parameters on Classification Algorithms
- To get a Best Performance of Classification Algorithm

#### SCOPE

- Develop a Machine Learning for business growing
- Reduce human mistaken
- Productivity and Efficiency in planning management
- Reduce noises to prevent bad learning to Machine Learning model

#### RESEARCH METHODOLOGY



#### ALGORITHMS USED

- Logistic Regression
- Decision Tree
- Random Forest
- SVM



#### ANALYSIS

| Algorithms          | Training | Fine-Tuning | Test |
|---------------------|----------|-------------|------|
| Logistic Regression | 71%      | 85%         | 86%  |
| Decision Tree       | 94%      | 80%         | 81%  |
| Random Forest       | 94%      | 90%         | 90%  |
| SVM                 | 66%      | 80%         | 85%  |

In the result, the Random Forest is a higher performance on this research project. In training model, fine-tuning and test model shown the Random Forest is higher accurate to classify the project remain budget as a determine the project profit.

#### CONCLUSION

This research project is success developed a Machine Learning model with Random Forest. The Random Forest is a best performance selected to developed in the Machine Learning mode. It is success carry the Artificial Intelligence benefit into the business field. This research project success prove the Machine Learning model increase the efficiency and productivity in decision making.

#### Recommendations

- Improve the dataset quality and more meaningful from the construction environment effective.
- Apply Deep Learning algorithms in the Machine Learning
- Explore more literature review to understand the Machine Learning model in construction planning.

# PLAGIARISM CHECK RESULT

## Machine Learning for data classification in construction project planning

### ORIGINALITY REPORT

|                  |                  |              |                |
|------------------|------------------|--------------|----------------|
| <b>7%</b>        | <b>6%</b>        | <b>3%</b>    | <b>3%</b>      |
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

### PRIMARY SOURCES

|          |   |               |
|----------|---|---------------|
| <b>1</b> | <b>data.cityofnewyork.us</b><br>Internet Source                       | <b>1%</b>     |
| <b>2</b> | <b>Submitted to Liverpool John Moores University</b><br>Student Paper | <b>&lt;1%</b> |
| <b>3</b> | <b>www.jestr.org</b><br>Internet Source                               | <b>&lt;1%</b> |
| <b>4</b> | <b>doctorpenguin.com</b><br>Internet Source                           | <b>&lt;1%</b> |
| <b>5</b> | <b>www.mdpi.com</b><br>Internet Source                                | <b>&lt;1%</b> |
| <b>6</b> | <b>managementpapers.polsl.pl</b><br>Internet Source                   | <b>&lt;1%</b> |
| <b>7</b> | <b>eprints.utar.edu.my</b><br>Internet Source                         | <b>&lt;1%</b> |
| <b>8</b> | <b>uobrep.aws.openrepository.com</b><br>Internet Source               | <b>&lt;1%</b> |

|    |   |      |
|----|---|------|
| 9  | Ching-Lung Fan. "Evaluation of Classification for Project Features with Machine Learning Algorithms", Symmetry, 2022<br>Publication   | <1 % |
| 10 | Submitted to Mantis College<br>Student Paper  | <1 % |
| 11 | upcommons.upc.edu<br>Internet Source  | <1 % |
| 12 | Submitted to University of Greenwich<br>Student Paper   | <1 % |
| 13 | pt.scribd.com<br>Internet Source  | <1 % |
| 14 | Yanan Yin, Fengtao Liu, Kai Li, Subei Tan et al. "Integrated proteomics analysis in cerebrospinal fluid and saliva reveals the changes of endopeptidase activity in Parkinson's disease", Research Square Platform LLC, 2023<br>Publication | <1 % |
| 15 | naco.gov.in<br>Internet Source  | <1 % |
| 16 | Submitted to University of Computer Studies<br>Student Paper  | <1 % |
| 17 | Atik Mahabub. "A robust voting approach for diabetes prediction using traditional machine   | <1 % |



learning techniques", SN Applied Sciences,  
2019  
Publication

---

|    |   |      |
|----|---|------|
| 18 | <a href="http://www.jescae.com">www.jescae.com</a><br>Internet Source   | <1 % |
| 19 | YeEun Jang, JeongWook Son, June-Seong Yi.<br>"Classifying the Level of Bid Price Volatility<br>Based on Machine Learning with Parameters<br>from Bid Documents as Risk Factors",<br>Sustainability, 2021<br>Publication | <1 % |
| 20 | Submitted to University of Northumbria at<br>Newcastle<br>Student Paper   | <1 % |
| 21 | <a href="http://www.hindawi.com">www.hindawi.com</a><br>Internet Source   | <1 % |
| 22 | <a href="http://ir.lib.uwo.ca">ir.lib.uwo.ca</a><br>Internet Source   | <1 % |
| 23 | <a href="http://www.jstage.jst.go.jp">www.jstage.jst.go.jp</a><br>Internet Source   | <1 % |
| 24 | Sophia Daskalaki, Ioannis Kopanas, Nikolaos<br>Avouris. "EVALUATION OF CLASSIFIERS FOR<br>AN UNEVEN CLASS DISTRIBUTION<br>PROBLEM", Applied Artificial Intelligence,<br>2006<br>Publication                             | <1 % |

---

|    |  |      |
|----|--|------|
| 25 | <a href="http://epubs.scu.edu.au">epubs.scu.edu.au</a><br>Internet Source  | <1 % |
| 26 | Santhosha Kamath, Apoorv Srivastava, Prathamesh Kamath, Sanjay Singh, M.Sathish Kumar. "Application Aware Multiple Constraint Optimal Paths for Transport Network using SDN", IEEE Transactions on Network and Service Management, 2021<br>Publication | <1 % |
| 27 | Submitted to University of Sunderland<br>Student Paper   | <1 % |
| 28 | <a href="http://www.ofzenandcomputing.com">www.ofzenandcomputing.com</a><br>Internet Source  | <1 % |
| 29 | Submitted to La Trobe University<br>Student Paper  | <1 % |
| 30 | <a href="http://www.hivmr.com">www.hivmr.com</a><br>Internet Source  | <1 % |
| 31 | <a href="http://link.springer.com">link.springer.com</a><br>Internet Source  | <1 % |
| 32 | <a href="http://moam.info">moam.info</a><br>Internet Source  | <1 % |
| 33 | <a href="http://vbn.aau.dk">vbn.aau.dk</a><br>Internet Source  | <1 % |
| 34 | <a href="http://fict.utar.edu.my">fict.utar.edu.my</a><br>Internet Source  | <1 % |

|    |  |     |
|----|--|-----|
| 35 | stce.huce.edu.vn<br>Internet Source  | <1% |
| 36 | businessdocbox.com<br>Internet Source  | <1% |
| 37 | pure.ewha.ac.kr<br>Internet Source   | <1% |
| 38 | www.jait.us<br>Internet Source   | <1% |
| 39 | www.mecs-press.org<br>Internet Source  | <1% |
| 40 | Florence Ogonjo, Angeline Wairegi, Joseph Gitonga. "Leveraging AI in the Kenyan Judiciary: A Case for Utilizing Text Classification Models for Data Completeness in Case Law Meta Data in Kenya's Employment and Labor Relations Court", Research Square Platform LLC, 2023<br>Publication | <1% |
| 41 | academic.oup.com<br>Internet Source  | <1% |
| 42 | catalog.ihsn.org<br>Internet Source  | <1% |
| 43 | civil.utm.my<br>Internet Source  | <1% |
| 44 | ebin.pub<br>Internet Source  |     |

|    |   |     |
|----|---|-----|
|    |   | <1% |
| 45 | ejournal.uniks.ac.id<br>Internet Source | <1% |
| 46 | eprajournals.com<br>Internet Source     | <1% |
| 47 | www.arxiv-vanity.com<br>Internet Source | <1% |
| 48 | www.diva-portal.se<br>Internet Source   | <1% |
| 49 | www.researchgate.net<br>Internet Source | <1% |

Exclude quotes On      Exclude matches < 8 words  
Exclude bibliography On

|  |            |                 |                  |
|--|------------|-----------------|------------------|
| <b>Universiti Tunku Abdul Rahman</b>   |            |                 |                  |
| <b>Form Title : Supervisor's Comments on Originality Report Generated by Turnitin for Submission of Final Year Project Report (for Undergraduate Programmes)</b> |            |                 |                  |
| Form Number: FM-IAD-005  | Rev No.: 0 | Effective Date: | Page No.: 1 of 1 |




**FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY**

|                                     |   |
|-------------------------------------|---|
| <b>Full Name(s) of Candidate(s)</b> | LAU ZHENG LIANG   |
| <b>ID Number(s)</b>                 | 17ACB05620  |
| <b>Programme / Course</b>           | IB  |
| <b>Title of Final Year Project</b>  | Machine Learning for data classification in construction project planning |

| <b>Similarity</b>   | <b>Supervisor's Comments<br/>(Compulsory if parameters of originality exceeds the limits approved by UTAR)</b> |
|---|--|
| <b>Overall similarity index:</b> <u>7</u><br><br><b>% Similarity by source</b><br>Internet Sources: <u>6</u> %<br>Publications: <u>3</u> %<br>Student Papers: <u>3</u> %  |  |
| <b>Number of individual sources listed of more than 3% similarity:</b> <u>0</u>   |  |
| <b>Parameters of originality required and limits approved by UTAR are as Follows:</b><br>(i) Overall similarity index is 20% and below, and<br>(ii) Matching of individual sources listed must be less than 3% each, and<br>(iii) Matching texts in continuous block must not exceed 8 words<br><i>Note: Parameters (i) – (ii) shall exclude quotes, bibliography and text matches which are less than 8 words.</i> |  |

Note Supervisor/Candidate(s) is/are required to provide softcopy of full set of the originality report to Faculty/Institute

*Based on the above results, I hereby declare that I am satisfied with the originality of the Final Year Project Report submitted by my student(s) as named above.*

  
 \_\_\_\_\_  
 Signature of Supervisor  
 Name: Cik Zanariah Binti Zainudin  
 Date: 8 September 2023

\_\_\_\_\_  
 Signature of Co-Supervisor  
 Name: \_\_\_\_\_  
 Date: \_\_\_\_\_



**UNIVERSITI TUNKU ABDUL RAHMAN**

**FACULTY OF INFORMATION & COMMUNICATION  
TECHNOLOGY (KAMPAR CAMPUS)**

**CHECKLIST FOR FYP2 THESIS SUBMISSION**

|                 |                             |
|-----------------|-----------------------------|
| Student Id      | 17ACB05620                  |
| Student Name    | LAU ZHENG LIANG             |
| Supervisor Name | CIK ZANARIAH BINTI ZAINUDIN |

| TICK (✓) | DOCUMENT ITEMS  |
|----------|---|
|          | Your report must include all the items below. Put a tick on the left column after you have checked your report with respect to the corresponding item.            |
| -        | Front Plastic Cover (for hardcopy)  |
| ✓        | Title Page  |
| ✓        | Signed Report Status Declaration Form   |
| ✓        | Signed FYP Thesis Submission Form   |
| ✓        | Signed form of the Declaration of Originality   |
| ✓        | Acknowledgement   |
| ✓        | Abstract  |
| ✓        | Table of Contents   |
| ✓        | List of Figures (if applicable)   |
| ✓        | List of Tables (if applicable)  |
| -        | List of Symbols (if applicable)   |
| ✓        | List of Abbreviations (if applicable)   |
| ✓        | Chapters / Content  |
| ✓        | Bibliography (or References)  |
| ✓        | All references in bibliography are cited in the thesis, especially in the chapter of literature review  |
| -        | Appendices (if applicable)  |
| ✓        | Weekly Log  |
| ✓        | Poster  |
| ✓        | Signed Turnitin Report (Plagiarism Check Result - Form Number: FM-IAD-005)  |
| ✓        | I agree 5 marks will be deducted due to incorrect format, declare wrongly the ticked of these items, and/or any dispute happening for these items in this report. |

\*Include this form (checklist) in the thesis (Bind together as the last page)

I, the author, have checked and confirmed all the items listed in the table are included in my report.

(Signature of Student)

Date: 11 September 2023