**LINEAR REGRESSION AS GLASS-BOX MODEL: A STUDY OF WORK STATUS GRADUATE EMPLOYABILITY OF PUBLIC UNIVERSITIES IN MALAYSIA (2018-2021)**

**HANANEE BUERAHENG**

**632437001**

**FATONI UNIVERSITY**

**1444/2023**

**LINEAR REGRESSION AS GLASS-BOX MODEL: A STUDY OF WORK STATUS GRADUATE EMPLOYABILITY PUBLIC UNIVERSITIES IN MALAYSIA (2018-2021)**

**HANANEE BUERAHENG**

**632437001**

**A PROJECT REPORT SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE DEGREE OF BACHELOR OF SCIENCE (DATA SCIENCE AND ANALYTICS)**

**FATONI UNIVERSITY**

**1444/2023**

**DEPARTMENT OF DATA SCIENCE AND ANALYTICS**

**FACULTY OF SCIENCE AND TECHNOLOGY**

**FATONI UNIVERSITY**

**PROJECT APPROVAL**

**Title :**

**Presented by :** Hananee Bueraheng 632437001

**IT HAS BEEN PRESENTED ON**

**Date :** 22 December 2023

**Location :**

**AGREED TO**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Mr. Maroning Useng Dr. Nurul Izrin Bt Md Saleh

Advisor Co Advisor

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Mr. Abdunfatah Masamae Dr. Sahidan Abdulmana

Committees I Committees II

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Assist. Prof. Dr. Anuwat Worlee

Dean, Faculty of Science and Technology

**Title :**

**Author :** Hananee Bueraheng, 632437001

**Department :** Data Science and Analytics

**Academic Year :** 2023

**Advisor :** Mr. Maroning Useng

**Co-advisor :** Dr Nurul Izrin Bt Md Saleh

# **ABSTRACT**

# **ACKNOWLEDGEMENT**

In the name of Allah. The Most Gracious and The Most Merciful.

First of all, I would like to dedicate my appreciation to Allah, The Most Merciful and The Most Compassionate who h

My supervisor from Fatoni University, Mr. Maroning Useng, for his enthusiasm, patience, insightful comments, helpful information, practical advice, and tireless ideas helped me greatly throughout my research and writing this paper. I would like to express my heartfelt gratitude. Thanks to his vast knowledge, deep experience, and expertise in data quality management, I was able to complete this study. Without his support and guidance, this project would not have been possible, I couldn't imagine a better director for my research.

I’d like to thank my external supervisor from Universiti Teknikal Malaysia Melaka (Dr. Nurul Izrin Binti Md Saleh) who helped me learn a lot about this project. Her ideas and comments aided in the completion of this project.

I also wish to express my sincere thanks to the Fakulti Teknologi Maklumat dan Komunikasi (FTMK), Universiti Teknikal Malaysia Melaka accepting me into the cooperative education program. In addition, I am deeply indebted to the Universiti Teknikal Malaysia Melaka for granting me free accommodation and a comfortable research environment. This support has enabled me to complete my Bachelor's Degree studies successfully.

I am grateful to my lecturers from Fatoni University for providing me with such a significant change throughout 8 semesters. I believe I will participate in more such activities in the future. I guarantee that this project was created entirely by me and is not a forgery. Finally, I’d like to express my gratitude to my parents and friends for their excellent comments and guidance during the completion of this project.

Hananee Bueraheng

# **TABLE OF CONTENTS**

Table of Contents

[**ABSTRACT** III](#_Toc153883052)

[**ACKNOWLEDGEMENT** IV](#_Toc153883053)

[**TABLE OF CONTENTS** V](#_Toc153883054)

[**LIST OF TABLES** VII](#_Toc153883055)

[**LIST OF FIGURES** VIII](#_Toc153883056)

[**CHAPTER I** 1](#_Toc153883057)

[1.1 PROJECT OVERVIEW 1](#_Toc153883058)

[1.2 PROBLEM STATEMENT 1](#_Toc153883059)

[1.3 PROJECT OBJECTIVE 2](#_Toc153883060)

[1.4 SCOPE OF STUDY 2](#_Toc153883061)

[1.5 SIGNIFICANT OF STUDY 2](#_Toc153883062)

[1.6 CHAPTER ORGANIZE 3](#_Toc153883063)

[**CHAPTER II** 4](#_Toc153883064)

[2.1 DEFINITION 4](#_Toc153883065)

[2.1.1 Public University 4](#_Toc153883066)

[2.1.2 Graduate Employability 5](#_Toc153883067)

[2.1.3 Linear Regression 6](#_Toc153883068)

[2.1.4 Glass-box model 6](#_Toc153883069)

[2.2 REQUIRED TOOLS 8](#_Toc153883070)

[2.2.1 Github 8](#_Toc153883071)

[2.2.2 Programming Language 8](#_Toc153883072)

[2.2.3 Libraries 8](#_Toc153883073)

[2.3 RELATED WORKS 9](#_Toc153883074)

[2.3.1 Soft Skills and Graduate Employability: Evidence from Malaysian Tracer Study 9](#_Toc153883075)

[2.3.2 Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications 9](#_Toc153883076)

[2.3.3 Paper 3 10](#_Toc153883077)

[**CHAPTER III** 11](#_Toc153883078)

[3.1 Business Understanding 12](#_Toc153883079)

[3.2 Data Understanding 12](#_Toc153883080)

[3.3 Data Preparation 16](#_Toc153883081)

[3.4 Modeling 17](#_Toc153883082)

[3.5 Evaluation 18](#_Toc153883083)

[3.6 Deployment 18](#_Toc153883084)

[**CHAPTER IV** 19](#_Toc153883085)

[**CHAPTER V** 24](#_Toc153883086)

[References 25](#_Toc153883087)

# **LIST OF TABLES**

[Table 1: List of public universities in Malaysia [2] XIII](#_Toc153882192)

[Table 2: Independent variables in the study XIX](#_Toc153882193)

[Table 3: Percentage of graduate employability status XX](#_Toc153882194)

[Table 4: Descriptive statistics for categorical variables XX](#_Toc153882195)

# **LIST OF FIGURES**

[Figure 1: Best Fit Line for a Linear Regression Model [5] XIV](#_Toc153882200)

[Figure 2: CRISP-DM diagram XVIII](#_Toc153882201)

[Figure 3: Data preparation process XXII](#_Toc153882202)

[Figure 4: Handling missing and infinite values XXII](#_Toc153882203)

[Figure 5: Example of variable selection XXII](#_Toc153882204)

[Figure 6: Data splitting XXIII](#_Toc153882205)

[Figure 7: Example of building a linear regression model XXIII](#_Toc153882206)

[Figure 8: Example of model evaluation XXIV](#_Toc153882207)

[Figure 9: Result visualization of status working XXV](#_Toc153882208)

[Figure 10: Result in the visualization of status not working yet XXVI](#_Toc153882209)

[Figure 11: Result in the visualization of the status of further study XXVII](#_Toc153882210)

[Figure 12: Result in the visualization of status waiting for job placement XXVIII](#_Toc153882211)

[Figure 13: Result in the visualization of status improvement skill XXVIII](#_Toc153882212)

# **CHAPTER I**

**INTRODUCTION**

## PROJECT OVERVIEW

Higher education, providers, and other stakeholders are concerned about recent changes in graduate employment. The "Linear Regression as a Glass-Box Model: A Study of Work Status and Graduate Employability in Public Universities in Malaysia (2018-2021)" project aims to investigate the factors influencing the work status of graduates from Malaysian public universities during the specified timeframe. Recognizing the significance of transparent and interpretable models, this study employs linear regression as a "glass-box" model to analyze the relationship between employment outcomes and nationality, field of study, and other relevant variables more transparent and understandable. The characteristics of “glass-box” model through the linear regression analysis allows (i) a simple identification of the overall factors of graduate employability, and, for each individual prediction, (ii) the attribution to each of the control variables—and/ or to their potential interactions—of the result obtained. These distinctive characteristics have important implications for comprehensive understanding of graduate careers, social inequality within the labor market, and the role of educational credentials in facilitating access to desired employment. [glass-box]

## PROBLEM STATEMENT

The graduate employability rate is now a big topic in the discussion among university boards because it is a ~~measure~~ benchmark of a university's success in producing many graduates. The high number of graduates is not proof of the success of producing human capital and becoming a student.? But what is the fate of the graduates after they finish their studies at the university that needs to be discussed. Employability features more prominently on the agenda of higher education institutions when the economy falters or changes: the majority of students, and their families [1].

Despite the increasing emphasis on higher education and the important role of public universities in shaping the future workforce, there remains a need to understand the determinants of graduate employability in Malaysia fully. The global and national economies are undergoing dynamic transitions between 2018 and 2021, which will have an impact on university graduates' career paths and the labor market. Based on this background, this research study aims to investigate the graduate employability work status of public university graduates over a certain period (2018-2021), with a particular focus on factors that influence graduate employability.

In terms of this, the purpose of this ~~research~~ study is to look into the employment situation of graduates from Malaysian public universities within the given period, with an emphasis on the variables affecting graduate employability. By using linear regression as a "glass-box" model, ~~researchers want~~ this study aims to make the complex relationship between employment outcomes and nationality, field of study, and other relevant variables more transparent and understandable.

## PROJECT OBJECTIVE

The purpose of this research study is to use linear regression modeling, to have a thorough grasp of graduate employability in Malaysian public universities from 2018 to 2021. The main objectives of this study are:

* Involve projecting the percentage of graduates who will enter the workforce by modeling their job status - ….. Furthermore, the study looks into what influences graduates' choices to continue their education, and develop a ~~solid~~ linear regression model to forecast the probability of pursuing these kinds of academic study. Examine waiting times for post-graduation job placements, using linear regression to identify significant variables and evaluate how well the model captures variability in waiting times.
* Assessing how accurate these projections are. assess graduates' aspirations to improve their skills and create a predictive linear regression model. Assess the model's capacity to pinpoint variables influencing graduates' drive to improve their skills.

## SCOPE OF STUDY

This study was conducted on graduates from 20 public universities in Malaysia who graduated in the year 2018-2021. The scope of this study includes an analysis of the employability of graduates of public universities in Malaysia from 2018 to 2021.

The survey focuses on key aspects such as employment status, intention to study further, waiting time for employment, and graduates' motivation to improve their skills. This study uses linear regression models to investigate the predictive relationships between various demographic and educational variables and the aforementioned dimensions of employability. Demographic factors such as gender, nationality, and educational background are investigated to understand the impact they have on graduate performance.

This study aims to provide a nuanced understanding of the factors that influence employability and provide useful insights for policies and educational strategies aimed at optimizing graduate career outcomes in Malaysian higher education.

## SIGNIFICANT OF STUDY

This study can be used by government and private agencies, economists, academics, and individuals interested in planning, policy formulation, analysis, and projection. It can help plan the development of the country's higher education.

* Policy formulation: Policymakers can use research findings to develop evidence-based policies aimed at improving graduate outcomes. Insights gained from linear regression models will be incorporated into targeted interventions to improve employment rates, reduce job wait times, and promote a culture of continuous skill development.
* University planning: This study provides strategic insights for public universities in Malaysia and provides a nuanced understanding of the factors that influence graduate employability. This knowledge helps universities adapt their academic programs, career services, and skills development initiatives to industry needs.

## CHAPTER ORGANIZE

**Chapter I: Introduction**

This chapter introduces about overview of the project, problem statement, objectives to analyze the graduate employability of the graduates from public universities from 2018-2021, project objective, scope of study, and software and hardware requirements.

**Chapter II: Literature View**

This chapter explains the theories from many sources such as articles journals, publications, conference proceedings, books, or case studies related to this project to be benefits and guidelines in handling this project.

**Chapter III: Methodology**

This chapter discusses the methodology used in this research project. It is the most crucial part because it will describe each phase of the project process in detail. For this project, the method used is the CRISP-DM framework.

**Chapter IV: Result**

In this chapter, I present and analyze the results of the linear regression models developed to predict various aspects of graduate employability in Malaysian public universities.

**Chapter V: Conclusion and Discussion**

In this chapter, I present and analyze the results of the linear regression models developed to predict various aspects of graduate employability in Malaysian public universities.

# **CHAPTER II**

**LITERATURE VIEW**

## DEFINITION

## Public University

Public universities or *Universiti Awam* in Malaysia can be categorized into 3 categories: research universities, focused universities, and comprehensive universities. Until 2023, 20 institutions of public universities provide educational services in Malaysia.

|  |  |
| --- | --- |
| Research Universities | Universiti Malaya (UM) |
| Universiti Kebangsaan Malaysia (UKM) |
| Universiti Teknologi Malaysia (UTM) |
| Universiti Sains Malaysia (USM) |
| Universiti Putra Malaysia (UPM) |
| Comprehensive Universities | Universiti Teknologi MARA (UiTM) |
| Universiti Islam Antarabangsa Malaysia (UIAM) |
| Universiti Malaysia Sabah (UMS) |
| Universiti Malaysia Sarawak (UNIMAS) |
| Focused Universities | Universiti Utara Malaysia (UUM) |
| Universiti Pendidikan Sultan Idris (UPSI) |
| Universiti Tun Hussein Onn Malaysia (UTHM) |
| Universiti Teknikal Malaysia Melaka (UTeM) |
| Universiti Malaysia Perlis (UniMAP) |
| Universiti Malaysia Terengganu (UMT) |
| Universiti Malaysia Pahang (UMP) |
| Universiti Sains Islam Malaysia (USIM) |
| Universiti Sultan Zainal Abidin (UnisZA) |
| Universiti Malaysia Kelantan (UMK) |
| Universiti Pertahanan Nasional Malaysia (UPNM) |

Table 1: List of public universities in Malaysia [2]

## Graduate Employability

According to the Ministry of Higher Education Malaysia, the graduate employability rate of a graduate is assessed from 5 statuses whether working, continuing their studies, improving their skills, waiting for job placement, or not working within 6 months after they have completed their studies [3]. Graduate employability can also be defined as the capacity of an individual who has finished their education to join, stay in, and grow within the workforce [4].

Within the framework of my research study, graduate employability can be defined as the effective transfer of graduates from school to the workforce, taking into account various criteria such as their study fields, employment status, and motivations for making career-related decisions.?

## Linear Regression

One type of supervised machine learning algorithm that computes a linear relationship between a dependent variable and one or more independent features is called linear regression. The independent and dependent variables are linearly related to each other. This means that the change in the dependent variable follows the change in the independent variable linearly. This means that there must be a straight line that can pass through the data points [5]. Below is an example of the best-fit line for a linear regression model.

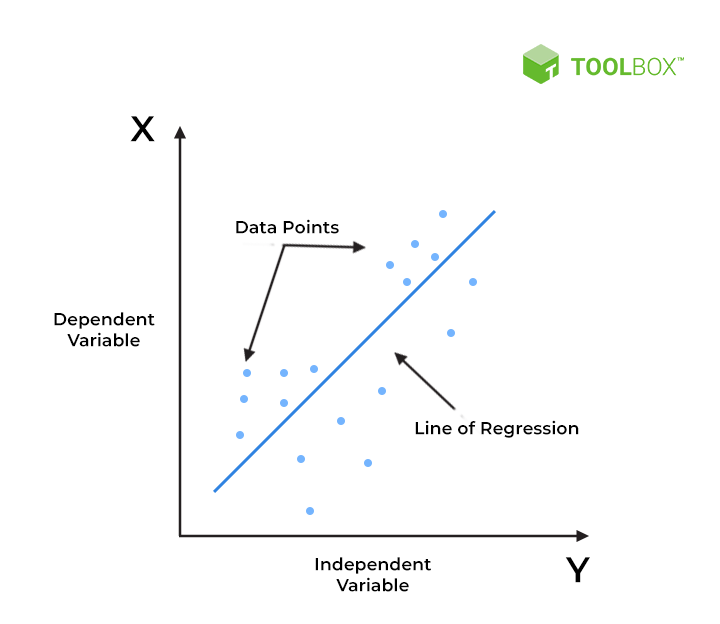


Figure : Best Fit Line for a Linear Regression Model [5]

In the above figure 1,

* X-axis = Independent variable
* Y-axis = Output / dependent variable
* Line of regression = Best fit line for a model

## Glass-box model

The glass box model is transparent to the user. In the glass box model, all features and model parameters are known to the user. We can also see the criteria the model uses to arrive at its predicted results and conclusions. This ensures complete transparency. Glass box models are more robust and easier to adapt to production environments [6].

Linear regression is considered a glass-box model based on several factors. Firstly, easy implementation. Linear regression models are computationally easy to implement because they do not require much technical effort before deployment or during model maintenance. It’s not the same as other deep learning models (neural networks), linear regression is relatively simple and it is straightforward. This makes the algorithm preferable to black box models that cannot justify which input variables cause changes in the output variables [5].

## REQUIRED TOOLS

## Github

The code hosting platform for version control and collaboration that I used throughout this study is Github. This allows me and others to collaborate on projects from anywhere. GitHub lets me create, save, modify, merge, and collaborate on files and code. Any member of the team can access her GitHub repository (think of it as a folder of files) and see the latest versions in real time [7]. I can then make edits and changes that my supervisor will see.

## Programming Language

* R language

## Libraries

* Tidyverse
* Caret

## RELATED WORKS

## Soft Skills and Graduate Employability: Evidence from Malaysian Tracer Study

Recent trends in the employment of graduate students have raised concerns among higher education providers and other stakeholders. As issues related to the soft skills gap have been raised repeatedly, industry and universities can take various initiatives to address this situation. Furthermore, as employment shifts from manufacturing to service industries, soft skills become more important. The purpose of this study is to map the soft skills and employment status of graduates after graduation. Using the Ministry of Higher Education's database, 100,413 first-year graduates who completed their studies were selected. Through logistic regression analysis, a graduate employability prediction model was constructed, with an accuracy of 77 degrees. Results show that 85.5% of the graduates were employed during data collection. The predictive model shows that graduates' employability status is influenced by factors such as gender, household income, field of study, MUET, CGPA, internships, entrepreneurship courses, work experience, communication skills, analytical skills, teamwork, and positive values [8].

## Counterfactuals and Causability in Explainable Artificial Intelligence: Theory, Algorithms, and Applications

This paper explores the challenge of comprehending and elucidating the predictions made by learning models used in various domains such, as medicine and autonomous vehicles. These models often pose difficulties for humans to understand, making it challenging to explain the reasoning behind their predictions.

To tackle this problem there has been an increasing interest in approaches that aim to enhance transparency and explainability in deep learning models. Some researchers argue that for a machine to be genuinely explainable to humans it should provide explanations that reveal cause-and-effect relationships that are causally comprehensible. Counterfactuals, a category of algorithms have the potential to offer causality.

The paper presents an analysis of existing literature on counterfactuals and causality within the context of artificial intelligence (AI). The authors employed Latent Dirichlet Topic Modeling (LDA) analysis within a Preferred Reporting Items, for Systematic Reviews and Meta-Analyses (PRISMA) framework to identify articles. This analysis resulted in the development of a taxonomy that considers the foundations of the examined algorithms, their characteristics, and real-world applications.

The research findings suggest that current model-agnostic counterfactual algorithms for explainable AI lack a strong grounding in causal theory. As a result, they do not effectively promote causability for human decision-makers. Moreover, popular algorithms in the literature tend to provide explanations based on spurious correlations rather than genuine cause-and-effect relationships. This can lead to suboptimal, incorrect, or biased explanations.

In conclusion, the paper highlights the need for more robust and theoretically grounded approaches to achieving causability in model-agnostic methods for explainable AI. It raises important challenges and directions for future research in this area [9].

## Paper 3

# **CHAPTER III**

**METHODOLOGY**

This chapter discusses the methodology used in this research project. It is the most crucial part because it will describe each phase of the project process in detail. For this project, the method used is the CRISP-DM framework.

****

Figure 2: CRISP-DM diagram

The CRISP-DM framework serves as a light of guidance in the field of data-driven research, offering a systematic way to extract insightful information from data. The Cross-Industry Standard Process for Data Mining, or CRISP-DM, provides a systematic and clear approach to handling challenging data analysis assignments.

At its core, CRISP-DM divides the data analysis process into six distinct phases:

## Business Understanding

Chapter I (Introduction) has explained the whole aspect that is required in the first phase of the CRISP-DM framework.

## Data Understanding

In this part, the study focuses on identifying, collecting, and analyzing the dataset that helps accomplish the project objectives.

**3.2.1 Collect data**

The data for this research study is secondary data downloaded from the Ministry of Higher Education official portal, page of *Laporan Kajian Pengesanan Graduan.* The data collection system, we called *Sistem Kajian Pengesanan Graduan (SKPG)* began to be collected two to three weeks before the date of convocation and closed one week after the end date of convocation every year. The dataset was downloaded from 2018-2021 as needed for this research study.

**Describe data**

The collected data selected as material for this research study is for 4 years starting from 2018-2021. These 4 years represent 3 important global phases that are important to understand the trends of graduate employability of graduates in Malaysia.

1. Phase 1 (2018-2019) shows the trends of graduate employability of graduates before the arrival of COVID-19.
2. Phase 2 (2020) shows the trends of graduate employability of graduates during the COVID-19 crisis.
3. Phase 3 (2021) shows the trends of graduate employability of graduates after the COVID-19 pandemic.

About Dataset

The dataset contains 2 parts; factors that influence graduate employability as independent variables and the patterns of work status of graduate employability as dependant variables. Table 2 lists all the variables used in the analysis.

|  |  |  |
| --- | --- | --- |
| Variable | Classification | Description |
| Gender | 1. Male 2. Female | Gender of the graduate |
| Nationality | 1. Malaysian 2. Non-Malaysian | Nationality of graduate |
| Level of study | 1. Diploma 2. First Degree 3. Master Degree 4. PhD 5. Another level | The highest level of study for graduate |
| Field of study | 1. Social science, business, and law 2. Engineering, manufacturing, and construction 3. Science, mathematics, and computers 4. Literature and humanities 5. Education 6. Health and welfare 7. Service 8. Agriculture and veterinary | Graduate’s field of study |

Table : Independent variables in the study

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | % | | | |
| 2018 | 2019 | 2020 | 2021 |
| Working | 58.6 | 63.5 | 60 | 61.6 |
| Not working yet | 19.8 | 13.8 | 15.6 | 14.5 |
| Further study | 15.7 | 16.7 | 17.8 | 18.8 |
| Waiting for job placement | 4.6 | 4.7 | 4.6 | 3.3 |
| Improve skill | 1.3 | 1.3 | 2 | 1.8 |

Table : Percentage of graduate employability status

As shown in Table 3, most fresh graduates were employed (working) about 6 months after they finished their studies.

Table 4 summarizes the demographics of the graduates in the 2018-2021 Graduate Tracer Study.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | % | | | |
| 2018 | 2019 | 2020 | 2021 |
| Gender |  |  |  |  |
| Male | 35.5 | 36.4 | 36.1 | 36.5 |
| Female | 64.5 | 63.6 | 63.9 | 63.5 |
|  |  |  |  |  |
| Nationality |  |  |  |  |
| Malaysian | 96.1 | 96 | 96.2 | 96 |
| Non-Malaysian | 3.9 | 4 | 3.8 | 4 |
|  |  |  |  |  |
| Level of study |  |  |  |  |
| Diploma | 21.7 | 21.9 | 24.2 | 25 |
| First Degree | 61.5 | 60.9 | 60.2 | 59.3 |
| Master Degree | 12.7 | 12.9 | 11.4 | 11.8 |
| PhD | 3.4 | 3.3 | 3 | 3.5 |
| Another level | 0.7 | 1.1 | 1 | 1.1 |
|  |  |  |  |  |
| Field of study |  |  |  |  |
| Social science, business, and law | 33 | 32.3 | 33.5 | 34.2 |
| Engineering, manufacturing, and construction | 22.9 | 24.8 | 23.4 | 22.1 |
| Science, mathematics, and computers | 15.9 | 14.1 | 15.4 | 15.2 |
| Literature and humanities | 9.6 | 9.7 | 9.9 | 9.6 |
| Education | 6.2 | 6.2 | 5.7 | 6.7 |
| Health and welfare | 5.8 | 5.9 | 5.1 | 5.6 |
| Service | 3.7 | 4 | 4.3 | 4.2 |
| Agriculture and veterinary | 0 | 3 | 2.7 | 2.2 |

Table : Descriptive statistics for categorical variables

In 2018, male graduates accounted for 35.5 of the study participants, while female graduates accounted for 64.5%. Regarding nationality, almost all 96.1% of the graduates are Malaysian and the rest are non-Malaysian. Regarding the level of study, more than half of graduates (61.5%) finished their study as First-Degree holder while 21.7% is Diploma holder, 12.7% is Master's Degree holder and 3.4% completed their study until Ph.D. and the rest 0.7% finished their study on another level. In the context of the field of study, a quarter (33%) received studies in social sciences, business and law. 22.9% of them studied engineering, manufacturing, and construction while 15.9% studied science, mathematics, and computers. Other than that, 9.6% of graduates finished their studies in literature and humanities meanwhile 6.2% were involved in the education field. A small amount 5.8% showed their interest in studies in health and welfare and the rest 3.7% in the field of services.

In 2019, male graduates accounted for 36.4 of the study participants, while female graduates accounted for 63.6%. Regarding nationality, almost all 96% of the graduates are Malaysian and the rest (4%) are non-Malaysian. Regarding the level of study, more than half of graduates (60.9%) finished their studies as First-Degree holders while 21.9% were Diploma holders, 12.9% were Master's Degree holders and 3.3% completed their studies as Ph.D. and the rest 1.1% finished their study on another level. In the context of the field of study, a quarter (32.2%) received studies in social sciences, business and law. 24.8% of them studied engineering, manufacturing, and construction while 14.1% studied science, mathematics, and computers. Other than that, 9.7% of graduates finished their studies in literature and humanities meanwhile 6.2% were involved in the education field. A small amount of 5.9% showed their interest in studies in health and welfare and 4% in the field of services. The rest 3% completed their studies in agriculture and veterinary.

In 2020, male graduates accounted for 36.1 of the study participants, while female graduates accounted for 63.9%. Regarding nationality, almost all 96.2% of the graduates are Malaysian and the rest (3.8%) are non-Malaysian. Regarding the level of study, more than half of graduates (60.2%) finished their study as First-Degree holders while 24.2% is Diploma holder, 11.4% is Master's Degree holder and 3% completed their study as PhD and the rest 1% finished their study on another level. In the context of the field of study, a quarter (33.5%) received studies in social sciences, business, and law. 23.4% of them studied engineering, manufacturing, and construction while 15.4% studied science, mathematics, and computers. Other than that, 9.9% of graduates finished their studies in literature and humanities meanwhile 5.7% were involved in the education field. A small amount of 5.1% showed their interest in studies in health and welfare and 4.3% in the field of services. The rest 2.7% completed their studies in agriculture and veterinary.

In 2021, male graduates accounted for 36.5 of the study participants, while female graduates accounted for 63.5%. Regarding nationality, almost all 96% of the graduates are Malaysian and the rest (4%) are non-Malaysian. Regarding the level of study, more than half of graduates (59.3%) finished their studies as First-Degree holders while 25% were Diploma holders, 11.8% were Master's Degree holders and 3.5% completed their studies until Ph.D. and the rest 1.1% finished their study on another level. In the context of the field of study, a quarter (34.2%) received studies in social sciences, business and law. 22.1% of them studied engineering, manufacturing, and construction while 15.2% studied science, mathematics, and computers. Other than that, 9.6% of graduates finished their studies in literature and humanities meanwhile 6.7% were involved in the education field. A small amount of 5.6% showed their interest in studies in health and welfare and 4.2% in the field of services. The rest 2.2% completed their studies in agriculture and veterinary.

## Data Preparation

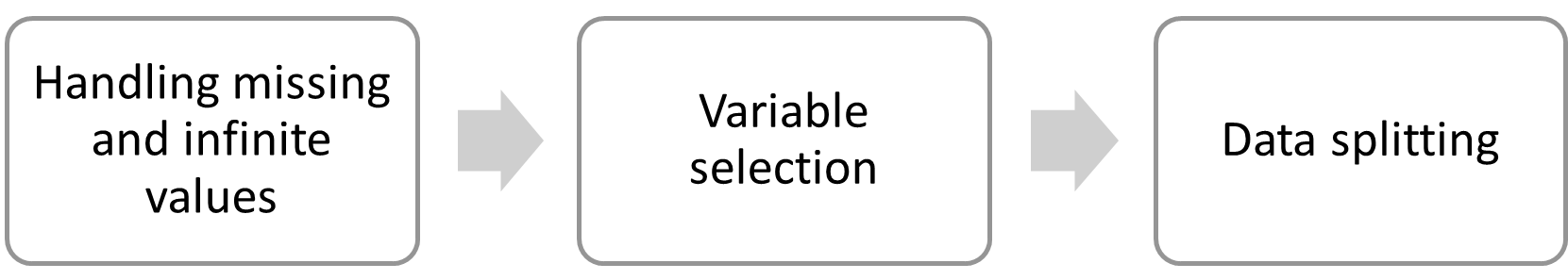


Figure 3: Data preparation process

* + 1. Handling missing and infinite values

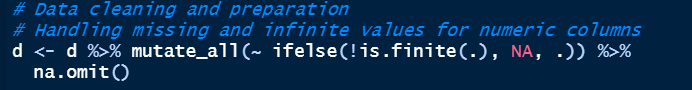


Figure : Handling missing and infinite values

In my context to encounter the missing and infinite values, I use 3 functions in 1 block of code:

* **mutate\_all**: Applies a function to all columns in the dataset.
* **~ ifelse(!is.finite(.), NA, .)**: Replaces non-finite values (infinite or NaN) with NA.
* **na.omit()**: Removes rows with any missing values.
  + 1. Variable selection

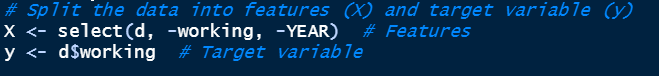


Figure : Example of variable selection

Variable selection is an important part of performing the model because it allows me to specify how independent variables are entered into the analysis. In the context of this research study, as mentioned in the data description above, independent and dependents have been clarified. We split the data into features (X) and target variables (y).

1. Feature matrix X: Independent variables
2. Target variables: Dependant variables
   * 1. Data splitting

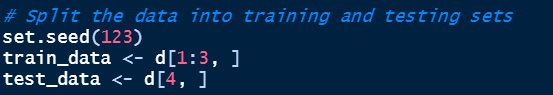


Figure : Data splitting

In this part, data splitting is dividing the dataset into training and testing sets. The training set is used to train the model, while the testing set is used to evaluate its performance. Typically, a common split is to allocate 80% of the data for training and 20% for testing because it represents the most variance of the dataset [10].

In this research study, we would split the data into two groups, the first three rows (2018-2020) assigned to the training set, and the fourth row (2021) assigned to the testing set. This partitioning ensures that the model is trained on a substantial portion of the data and evaluated on unseen data to assess its generalization ability.

## Modeling



Figure : Example of building a linear regression model

The linear regression model for predicting the percentage of graduates who are working is constructed using the following formula:

For example:

working∼predictors

* Variable: working

This is the dependent variable or target variable. It represents the percentage of graduates who are currently employed.

∼∼: In R syntax, this symbol denotes a formula that specifies the relationship between the dependent variable and the predictors.

predictors: This refers to the independent variables or features used to predict the target variable.

So, the formula essentially means that the percentage of graduates currently working is modeled as a linear combination of the selected predictors.

This linear regression model is then fitted to the training data, and predictions are made for the test data. The performance of the model is evaluated using metrics like RMSE, and the results are presented in the report. The actual vs. predicted values for the model are also displayed in a table within the report.

Explain Liner Regression and how it is Glass-Box Model

## Evaluation

The performance of each model was evaluated based on RMSE as shown in the table will be explained in Chapter 4. Root Mean Squared Error (RMSE) was chosen as the performance metric due to its suitability for regression tasks. The RMSE is calculated by taking the square root of the MSE.

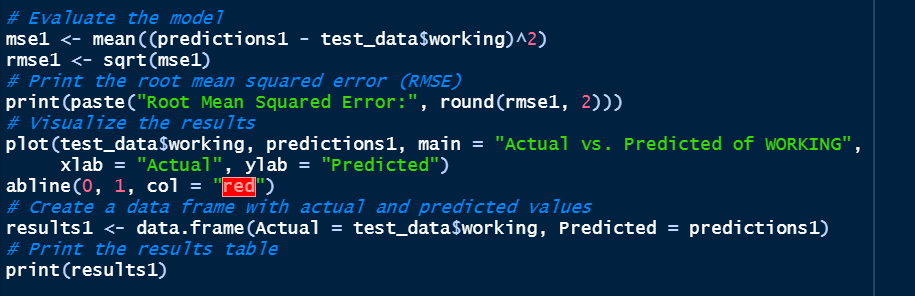


Figure : Example of model evaluation

This block of coding calculates the mean squared error (MSE) between the predicted values (predictions1) and the actual values of "working" in the test set (test\_data$working). Then, the RMSE is calculated by taking the square root of the MSE. RMSE is a measure of how well a model's predictions match actual values. The predictions for "working" vs the actual value of "working" are stored in the "results1" data frame.

The same steps are repeated to calculate the predicted values for other dependent variables (not working yet, further studies, waiting for job placement, and improving skills).

## Deployment

# **CHAPTER IV**

**RESULT AND INTERPRETATION**

In this chapter, I present and analyze the results of the linear regression models developed to predict various aspects of graduate employability in Malaysian public universities.

|  |  |  |  |
| --- | --- | --- | --- |
| Observed | % | | RSME |
| Actual | Predicted |
| Working | 61.6 | 61.4 | 0.4 |
| Not working yet | 14.5 | 12.85 | 1.65 |
| Further study | 18.8 | 18.84167 | 0.04 |
| Waiting for job placement | 3.3 | 4.616667 | 1.32 |
| Improve skill | 1.8 | 2.291667 | 0.49 |

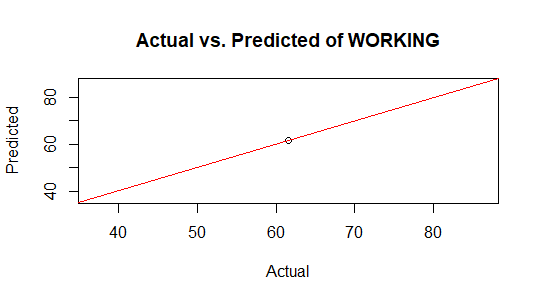
****

Figure : Result in the visualization of status working

Figure 9 visualizes the result of predicted vs value of percentage for status working. Based on the statistics dataset by the Malaysian Ministry of Higher Education, the actual value percentage of status working is 61.6%. However, the predicted values of the linear regression model were 61.4%. When using the model to predict "working" values on new data, we can expect the predicted values to deviate, on average, by approximately 0.2 units from the true values. Generally, a lower RSME value indicates a better fit of the model to the data. In this case, the RSME of 0.2 suggests that the "working" model provides relatively accurate predictions for the "working" variable.

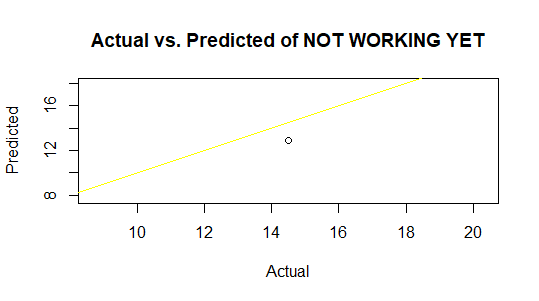
****

Figure : Result in the visualization of status not working yet

Figure 10 visualizes the result of predicted vs value of percentage for status not working yet. Based on the statistics dataset by the Malaysian Ministry of Higher Education, the actual value percentage of status working is 14.5%. However, the predicted values of the linear regression model were 12.85%. RMSE is a measure of how well a model's predictions match actual values. In this context, an RMSE of 1.65 indicates that the model's prediction of "not working yet" is, on average, within 1.65 units of the actual value in the test set.

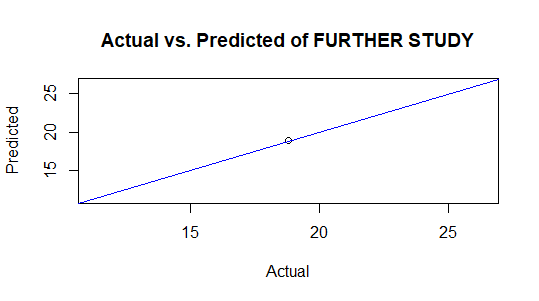
****

Figure : Result in the visualization of the status of further study

Figure 11 visualizes the result of predicted vs value of percentage for status further study. Based on the statistics dataset by the Malaysian Ministry of Higher Education, the actual values percentage of status further study is 18.8%. However, the predicted values of the linear regression model shows are 18.84167%. When using the model to predict "further study" values on new data, we can expect the predicted values to deviate, on average, by approximately 0.04 units from the true values. Generally, a lower RSME value indicates a better fit of the model to the data. In this case, the RSME of 0.04 suggests that the "further study" model provides relatively accurate predictions for the "further study" variable.

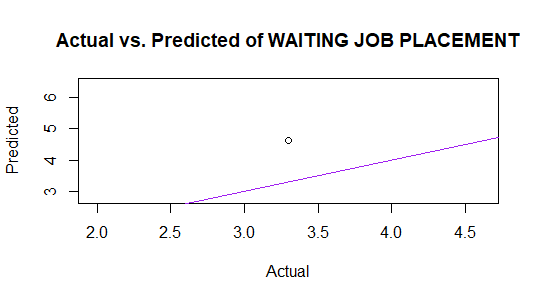
****

Figure : Result in the visualization of status waiting for job placement

Figure 11 visualizes the result of predicted vs value of percentage for status waiting for job placement. Based on the statistics dataset by the Malaysian Ministry of Higher Education, the actual value percentage of status waiting for job placement is 3.3%. However, the predicted values of the linear regression model shows are 4.616667%. RMSE is a measure of how well a model's predictions match actual values. In this context, an RMSE of 1.32 indicates that the model's prediction of " status waiting for job placement " is, on average, within 1.32 units of the actual value in the test set

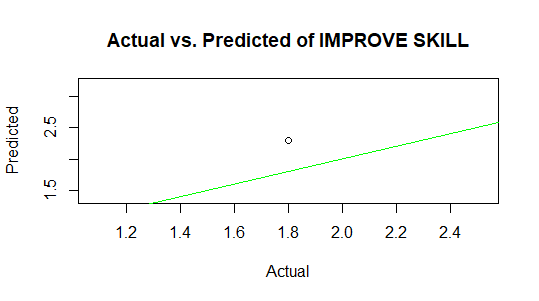
****

Figure : Result in the visualization of status improved skill

Figure 13 visualizes the result of predicted vs value of percentage for status improvement skill. Based on the statistics dataset by the Malaysian Ministry of Higher Education, the actual value percentage of status improvement skills is 1.8%. However, the predicted values of the linear regression model shows are 2.291667%. When using the model to predict "improve skill " values on new data, we can expect the predicted values to deviate, on average, by approximately 0.49 units from the true values. Generally, a lower RSME value indicates a better fit of the model to the data. In this case, the RSME of 0.49 suggests that the " improve skill " model provides relatively accurate predictions for the " improve skill " variable.

# **CHAPTER V**

**CONCLUSION**

All stakeholders, particularly higher education institutions and the Ministry of Higher Education must aggressively develop a strategic roadmap to increase future graduates’ readiness to align with labour market demand. This study provides insights into graduates’ employability concerning their field of study and their ability to acquire industry-required skills. The findings can aid in developing intervention programs to produce highly skilled workers, allowing fresh graduates to contribute to the country’s economic growth.

# References

|  |  |
| --- | --- |
| [1] | B. Oliver, "Redefining graduate employability and work-integrated learning: Proposals for effective higher education in disrupted economies," *Journal of Teaching and Learning for Graduate Employability,* vol. 5, no. 1, pp. 56-65, 2015. |
| [2] | M. o. H. Education, "Kategori Universiti Awam," Ministry of Higher Education, 18 January 2021. [Online]. Available: https://www.mohe.gov.my/en/institutions/public-university/kategori-ua. [Accessed 4 December 2023]. |
| [3] | M. o. H. Education, "GRADUATE EMPLOYABILITY; GE," Ministry of Higher Education, 2022. [Online]. Available: https://ge.mohe.gov.my/SearchGraduateEmployability.aspx. [Accessed 6 December 2023]. |
| [4] | D. M. Velliaris and D. Coleman-George, "Repositioning Study Abroad as a Rite of Passage: Impact, Implications, and Implementation," in *Handbook of Research on Study Abroad Programs and Outbound Mobility*, Hershey, Pennsylvania, USA, IGI Global, 2016, p. 26. |
| [5] | V. Kanade, "What Is Linear Regression? Types, Equation, Examples, and Best Practices for 2022," Spiceworks, 3 April 2023. [Online]. Available: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-linear-regression/. [Accessed 24 November 2023]. |
| [6] | C. Ndirutu, "Building a Glass-box model using InterprateML," 15 December 2021. [Online]. Available: https://www.section.io/engineering-education/how-to-build-a-glass-box-model-using-interpretml-python/. [Accessed 13 December 2023]. |
| [7] | C. Staff, "What Is GitHub and Why Should You Use It?," Coursera, 30 November 2023. [Online]. Available: https://www.coursera.org/articles/what-is-git. [Accessed 19 December 2023]. |
| [8] | N. M. Basir, Y. Z. Zubairi, R. Jani and D. Abdul Wahab, "Soft Skills and Graduate Employability: Evidence from Malaysian Tracer Study," *Pertanika Journal of Social Science and Humanities,* vol. 30, no. 4, pp. 1975-1989, 2022. |
| [9] | Y.-L. Chou, C. Moreira, P. D. Bruza and C. Ouyang, "Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications," *Information Fusion,* vol. 81, 2021. |
| [10] | A. A. Tokuç, "Splitting a Dataset into Train and Test Sets," Tarnum Java SRL, 12 May 2023. [Online]. Available: https://www.baeldung.com/cs/train-test-datasets-ratio. [Accessed 11 November 2023]. |