# WORKPLACE PREFERENCE ANALYTICS AMONG MMU GRADUATES

ONG SIN YIN

BACHELOR OF COMPUTER SCIENCE (HONS. DATA SCIENCE)

MULTIMEDIA UNIVERSITY

JANUARY 2023

# WORKPLACE PREFERENCE ANALYTICS AMONG MMU GRADUATES

BY

ONG SIN YIN

Bachelor of Computer Science (Hons. Data Science), Multimedia University, Malaysia

THESIS SUBMITTED IN FULFILMENT OF THE

REQUIREMENT FOR THE DEGREE OF

BACHELOR OF COMPUTER SCIENCE (HONS. DATA SCIENCE)

(by Research)

in the

FACULTY OF COMPUTING AND INFORMATICS

MULTIMEDIA UNIVERSITY
MALAYSIA

January 2023

## DECLARATION

I hereby declare that the work has been done by myself and no portion of the work contained in this thesis has been submitted in support of any application for any other degree or qualification of this or any other university or institute of learning.

_____

Ong Sin Yin
Faculty of Computing & Informatics
Multimedia University
Date: 09/02/2023

## ACKNOWLEDGEMENTS

It is a great pleasure to present this final year project with months of focused effort and dedication. I would like to take this opportunity to express my sincerest gratitude to those who have helped me along the way.

First and foremost, I would like to express my profound gratitude to my project supervisor Prof. Ts. Dr. Ting Choo Yee. He has been a great mentor who gives me invaluable guidance and advice during this project. His expertise and insight have helped to shape and strengthen my work significantly and eventually brought this project to its completion.

Secondly, I would like to acknowledge the support of the Faculty of Computing Informatics at Multimedia University, which provided the necessary resources and infrastructure that significantly assist me to carry out this research. This research would not have been feasible without my faculty's assistance.

Lastly, I am also grateful to my friends and classmates who have generously shared their knowledge with me and helped me to get a deeper understanding with beneficial information through discussions, even cheering me on throughout this journey and keeping me stay motivated to reach my targets.

Thank you all.

## ABSTRACT

This is a research project to study workplace preference among graduates by developing machine learning models on top of a web application. Since a number of graduates are struggling on being unemployment due to the issue of high competitiveness, loss of direction, and lack of skillset and experience, therefore a platform in helping them to find their desired workplace is high in demand nowadays to resolve the obstacles faced. In the meantime, it is important to find out how to determine the optimal factor to select a suitable job sector, how to identify the trustworthy data points for a better predictive model as well as which predictive models are most suited to be built to achieve the ultimate goal of this project.

This interim report covers the work done in the phase of FYP 1. By starting from project planning, an introduction of the project title will be interpreted with the objective, problem statement and scope of the project. Also, a literature review will be studied for referring to the proposed methods used. The flow of the project framework starts with performing data preparation on the main graduate dataset and 2 additional datasets which are location analytical datasets for graduates and companies. Next, data preprocessing will be carried out to improve the data quality. After that, a study of model construction and the strategies to evaluate the predictive model will ease the step of selecting the best-performing model for deployment sooner. Furthermore, a study of model construction and the strategies to evaluate the predictive model will ease the step of selecting the best-performing model for deployment sooner. Additionally, a finding will be conducted with data visualization for pattern discovery within the datasets. Lastly, the implementation plan of the next phase, FYP 2 and also conclusion will be outlined by the end of the report.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

A workplace is any space, where an employee permanently or temporarily performs any job-related task. Nevertheless, the workplace is also a place where university or college graduates will have to step in after they completed their studies. However, employment is always a challenge for university graduates who have insufficient employment skillsets and lack of work experience. It is common that university students put themselves aside from getting a job after graduation. As the number of graduates keeps rising, the employment situation of university graduates remains unoptimistic (Fan, 2020). Thus, the unemployment rate among recent graduates is rising which is an vital issue that universities should pay attention to (Megasari, Piantari, & Nugraha, 2020). According to tracer studies conducted by the Ministry of Higher Education, up to 25% of recent graduates in Malaysia remain without employment even six months after receiving their degrees, as stated by the government's Economic Planning Unit (Olowolayemo, Harun, & Mantoro, 2018).

Indeed, preparing students for employment is a significant mission of every university. The employment status of graduates reflects the quality of university education and the cultivation of talents. Even though universities and colleges are required by the Ministry of Education to report the employment status each year, an in-depth analysis still hasn't been well performed on the employment data (Yu & Zhang, 2018). Therefore, universities could take action with resources gathered from students in order to provide assistance for them to gain a decent job. A system that analyses the workplace preference of graduates would be a good solution to assist graduates through data mining approaches. A more streamlined hiring procedure would increase graduates' motivation and make it simpler for them to find work. Meanwhile, graduates would have a better idea on what's the next step they could plan ahead for their employment life. As a result, the findings of this research are expected to build a prediction model

that provides workplace recommendations to graduates to assist in their first steps in the workforce.

## 1.1 Problem Statement

Throughout the globe, the employment rate among graduates is dismally low. At the outset, it's important to note that many recent university graduates don't have a clear strategy for their future careers. Most of them have hazy notions of the job market and they haven't put in the time and effort necessary to thoroughly consider their job interests, personality, skills, beliefs, and goals for selecting a job (Jie, Zheng, Qi, & Xiya, 2021). Therefore, it is a vital problem to determine the optimal factors to select an appropriate job sector from numerous factors. There are various existing studies have identified factors to choose a career in both public and private sectors. For instance, work environment, promotion opportunities, salary, organisation culture, health safety and much more aspects (Rahman & Asadujjaman, 2021). These days, a graduate's preference on a career path is influenced by a wider range of criteria than ever before.

Secondly, in most cases, the career history of fresh graduates mostly will be blank hence they are short of work experience (Zhou, Liao, Ge, & Sun, 2019). Thus, this limits the available data that may be utilised, since it is only feasible to match student profiles with job requirements rather than past employee profiles. Industry and university both need to have an analysis to assess which graduates possess the knowledge and skills needed to succeed in their chosen fields. The profiles may not be a complete representation of a graduate's personality (Megasari et al., 2020). Therefore, it is challenging to determine the reliable data points that make predictive models perform well for finding their desired workplace based on the ability and conditions of a graduate in a relatively objective way.

Last but not least, the annually rising number of graduates has led to an expansion in the number of higher education institutions. Consequently, high competitive-

ness for job opportunities increases the pressure on new graduates (Li, Chuancheng, Hongguo, & Yanhui, 2018) while simultaneously slowing down the job application process due to the difficulty of sourcing skilled graduates that match the company's needs (Awujoola, 2021). Accordingly, there is an increasing demand for better methods of job matching for graduates. In this way, the issue worthy of dealing with is which predictive models are suitable to be developed in suggesting graduates an ideal workplace based on their preferences. Ultimately, the real-world problem raised above may be resolved by guiding graduates toward a sound solution.

## 1.2 Objectives

The objectives of this project are listed as below:

(i) To identify the optimal factors of selecting an appropriate job sector in helping graduates to make informed decisions.

(ii) To determine the reliable data points that make predictive models perform well in finding their desired workplace.

(iii) To find out which predictive models are suitable to be developed in suggesting graduates an ideal workplace based on graduates' profile obtained.

## 1.3 Project Scope

This is a research-based project to study workplace preference analytics among graduates. The data set utilised is real-world data provided by a Malaysian university to guarantee a reliable prediction model. This project will concentrate on Machine Learning models using Python programming language to deploy cluster analysis and classification models for predictive models. In addition, the front end of the web application that interacts with users and provides workplace suggestions to graduates is built using FlutterFlow.

## 1.4 Chapter Outline

In chapter 1, the general idea of this project topic is presented in depth with the problem statement, objectives and scope of the project.

Chapter 2 covers the literature review which presents the background study related to this project. The challenges faced by other researchers have to be investigated accordingly with the proposed methods and techniques from the researchers in order to find out how it relates to workplace preference analytics among graduates.

Besides, chapter 3 presents the methodology where the techniques chosen for implementation are discussed and explained in depth. The primary modelling strategies used are clustering and classification in this research.

Chapter 4 focuses on the preliminary data, along with a detailed description of the data and the process of data cleaning. In addition, data visualization presents the relationship discovered between the variables in the dataset.

Next, chapter 5 discusses the implementation plan which will be carried out in the phase of FYP2. A brief overview of the Gantt Chart, the target of the next phase and further implementation procedures are introduced in this chapter.

Lastly, Chapter 6 concludes the project by summarising its key findings, including what was accomplished and what remains to be done in the next phase of the project.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 Introduction

As higher education has getting improved in recent years, the rising number of graduates can be seen clearly but the employment status still remains unoptimistic. Numerous researchers have focused on the challenges of the low employment rate of graduates, with various methods and techniques to tackle particular problems. A good understanding of the causes and effects would be extremely beneficial for people to find a workplace for graduates.

## 2.2 Challenges faced by researchers

There are a number of factors that impact the need of constructing workplace preference analytics. Every researcher has to look into the challenges in order to develop a system that addresses the issue for graduates and industry.

The majority of researcher appear to be struggling with insufficient employment skills. Finding job vacancies is an obstacle for students who have recently completed their studies in higher education because they lack work experience hence they should look for job position that truly match their current criteria (Puspasari, Damayanti, Pramono, & Darmawan, 2021). Despite the presence of corresponding courses about teaching practice in the talent development programmes of related majors, students still lack some practical skills such as social skill and management skill as well as emergency handling (Jie et al., 2021). As a result, their criteria often do not meet the industry's expectations which resulting the failure of getting a job.

Aside from that, the following challenge that should be highlighted is employ-

ment dissatisfaction. A graduate typically finds it difficult to find a satisfactory job which meets his or her expectations when starting their first job, which eventually leads to poor performance and life dissatisfaction. Selection of an appropriate job sector can depend on the work environment, promotion opportunity, salary, increment of salary and incentives and many more (Rahman & Asadujjaman, 2021). On the other hand, industry is also confronted with the difficulty of recruiting skilled and satisfactory graduates that match their requirements (Awujoola, 2021). Employment dissatisfaction may lead to a drop in company efficiency and performance, which will have an impact on our economic growth of the country.

Due to the rapid development of the internet, the growth of data causes difficult aspects of handling and analysing it, even with recent advancements in computing resources (Rodriguez & Chavez, 2019). Employment data disorganised must be a vital problem for all organizations, especially universities and companies. Therefore, job matching becomes a difficult task for graduates as well as the companies.

Furthermore, most of the graduates are suffering from unclear employment planning. In general, graduates have no employment history. This seems to be having a weak employment concept, as well as a lack of clear and serious consideration and planning regarding their own preferences, characteristics, skillset and values, direction and goals (Jie et al., 2021).

Besides that, high competitiveness for a job opportunity can be seen nowadays. The number of higher education institutions has grown in accordance with the vastly increased number of graduates annually (Premalatha & Sujatha, 2021). Thus, a graduate needs to get through a keen competition among a huge amount of candidates to get a job which may increase the failure rate of the job application.

The least challenges that faced by the researchers is time consuming on job application. This is due to the fact that they have short of experience hence they spend their time screening a few jobs from hundreds of companies. After they have chosen

a few jobs to apply for, they must still go through the process of preparation and attending tests and interviews which are probably manually arranged by the company's employer (Zhou et al., 2019). Last but not least, the difficulties encountered by the researchers highlight the significance of developing the project's predictive model.

**Table 2.1: Challenges faced by researchers (Part 1)**

| Author | Unclear employment planning | Employment dissatisfaction | Insufficient employment skills | Employment data disorganised | High competitiveness for a job opportunity | Time consuming on job application |
|---|---|---|---|---|---|---|
| Jie et al. (2021) | x | | x | | | |
| Chen, Liang, Gao, Zhou, and Wu (2019) | | x | | | | |
| Fan (2020) | | x | | | | |
| Yu and Zhang (2018) | | | | x | | |
| Rahman and Asadujjaman (2021) | x | x | | | | |
| Nigam, Roy, Singh, and Waila (2019) | x | | | | | |
| Premalatha and Sujatha (2021) | | x | x | | | |
| Megasari et al. (2020) | | x | | | | |
| Zhou et al. (2019) | x | x | x | | | x |
| Nudin, Warsito, and Wibowo (2022) | | x | x | | | x |
| Kusnawi, Ipmawati, and Kusumandaru (2019) | | | x | | | |
| Chakraborty, Hossain, and Arefin (2019) | | | | | x | x |
| Olowolayemo et al. (2018) | | | | x | | x |
| Atela, Othuon, and Agak (2020) | x | | | | | |
| Rodriguez and Chavez (2019) | | | | x | | |
| Li, Chuancheng, Hongguo, and Yanhui (2018) | | x | | | | |
| Yang and Cao (2019) | | | | x | | |
| Puspasari et al. (2021) | | | x | | x | |
| Kumar, Prakash, and Anbuchelian (2020) | | | x | x | x | |
| Jiang (2019) | | | | x | x | |
| Hooshyar, Pedaste, and Yang (2020) | | | x | | | |
| Awujoola (2021) | | x | | x | | |

**Table 2.2: Challenges faced by researchers (Part 2)**

| Author | Unclear employment planning | Employment dissatisfaction | Insufficient employment skills | Employment data disorganised | High competitiveness for a job opportunity | Time consuming on job application |
|---|---|---|---|---|---|---|
| Nie et al. (2017) | x | | x | | | |
| Heriyadi (2021) | | x | | | | |
| Gunarathne, Senaratne, and Herath (2021) | | | x | | | |
| Kahn, Gamedze, and Oghenetega (2019) | | | x | | | |
| Belsare and Deshmukh (2018) | | x | | x | x | |
| Martinez-Gil, Freudenthaler, and Natschläger (2018) | | | | | x | x |
| Sharma, Joshi, Sharma, Singh, and Gupta (2021) | x | | | | | |
| Van Huynh, Van Nguyen, Nguyen, and Nguyen (2020) | x | x | | | x | |
| Vassef, Toosi, and Akhaee (2022) | | | | x | | |
| Zhu, Viaud, and Hudelot (2021) | | | | x | | |
| Uddin (2022) | x | | | x | x | |
| Li, Chuancheng, Yinggang, Hongguo, and Yanhui (2018) | | | | x | x | |
| Vinutha and Yogisha (2021) | | | | | x | |
| Bannaka, Dhanasekara, Sheena, Karunasena, and Pemadasa (2021) | x | | x | | | |
| Saeed, Sufian, Ali, and Rehman (2021) | | | | x | | |
| Almutairi and Hasanat (2018) | | | x | | | |
| Angesti, Kurniawati, and Anggana (2021) | | | x | | | x |
| Jamal et al. (2020) | x | | x | | | |
| Nachev and Teodosiev (2018) | | x | | x | | |
| Amalia1 and Wibowo (2020) | | | | | x | x |
| Arafath, Saifuzzaman, Ahmed, and Hossain (2018) | | | | x | x | |
| N. VidyaShreeram1 (2021) | | | | | | x |
| Ankush Daharwal1 (2020) | x | | | | | |
| Kamal, Naushad, Rafiq, and Tahzeeb (2021) | x | | | | | |
| Bharambe, Mored, Mulchandani, Shankarmani, and Shinde (2017) | | | | | | x |
| Sui, Chang, Hsiao, Chen, and Chen (2018) | | | x | | x | |
| D.Haritha (2019) | | x | | | | |
| Qin et al. (2018) | | x | | | | |

Table 2.1 and 2.2 shows the challenges and problems faced by researchers in achieving workplace preference analytics. Looking at the table, the problem that is faced the most is that student is struggling with insufficient employment skills which is 16 out of 50. The second most faced challenge is employment dissatisfaction, again numbers spoke, 15 out of 50 papers focused on the challenge of employment data disorganised as well. Meanwhile, the least faced challenges by the researches is Time consuming on job application, it having only 9 out 50.

## 2.3 Proposed methods by researchers

To overcome the challenges mentioned in Section 2.2, researchers have proposed a sound solution to construct workplace preference analytics. In this section, there are total of 10 proposed methods, each of which can be subdivided into a few more techniques. A thorough discovery is encouraged in order to gain a better understanding of how each method can help to address the issues and the differences.

The majority of the researchers have proposed a predictive model to solve the unemployment issue by predicting their workplace preferences. First of all, a classification model was proposed by most of the researchers. A classification algorithm is a supervised learning technique in data mining that makes predictions about the likelihood or probability that subsequent data will fall into one of the predetermined categories based on input data that has already been labelled. In other words, predictive modelling that predicts a class label for a given example of input data. For instance, classification is used for various filtering purposes, including determining whether an email is spam or not.

Besides, deep learning is the second approach of most researchers. Deep learning is a type of machine learning and artificial intelligence (AI) that attempts to simulate the way that humans learn new information. It is extremely beneficial to deal with the task of collecting, analysing, and interpreting massive amounts of data in a simplified and faster process. Computer vision, speech recognition, and medical

image analysis are just some of the applications that have benefited from the use of deep-learning architectures like deep neural networks, recurrent neural networks, convolutional neural networks, and others.

Furthermore, the third ranking method among those researchers are using ensemble learning method for workplace preference analytics. Ensemble learning is a machine learning paradigm in which multiple models, such as classifiers or experts, are generated and combined strategically to solve a specific computational intelligence problem. Ensemble methods are used in a wide variety of applications including but not limited to classification, regression, and control.

Next, clustering was placed as the fourth most proposed method. Clustering is an unsupervised learning method commonly used for finding collections of data that share commonalities. Cluster analysis is the process of segregating a set of data points into a number of groups/clusters based on their similar characteristics. In addition, similarity measures/correlation measures also have been proposed in order to assess how well the job requirements and the skillsets of the graduates match up.

Apart from that, the remaining methods are statistical data analysis, which seeks to discover patterns and relationships; regression model for prediction by estimating the relationships between a dependent variable and a set of independent variables; filtering model, which makes decisions based on similarities in features; association algorithm to find graduates' behaviour and dimensionality reduction, which used for data feature reduction.

**Table 2.3: Proposed methods by researchers (Part 1)**

| Author | Clustering | Classification | Regression | Association | Deep Learning | Ensemble Learning | Dimensionality Reduction | Filtering | Correlation/Similarity Measures | Statistical Data Analysis |
|---|---|---|---|---|---|---|---|---|---|---|
| Jie et al. (2021) | | | | | | | | | | |
| Chen et al. (2019) | x | | | | | | | | | |
| Fan (2020) | | x | | | | | | | | |
| Yu and Zhang (2018) | | x | | x | | | | | | |
| Rahman and Asadujjaman (2021) | x | | | | | | | | | |
| Nigam et al. (2019) | | | | | x | | | | | |
| Premalatha and Sujatha (2021) | | x | x | | x | x | | | | |
| Megasari et al. (2020) | x | | | | | | x | | | |
| Zhou et al. (2019) | x | | | | | | | x | | |
| Nudin et al. (2022) | | x | | | | | | | | |
| Kusnawi et al. (2019) | | | | | | | | | x | |
| Chakraborty et al. (2019) | | | | | | | | | x | |
| Olowolayemo et al. (2018) | | | | | | | | | x | |
| Atela et al. (2020) | | | | | | | | | | x |
| Rodriguez and Chavez (2019) | | | | | | | | | x | |
| Li, Chuancheng, Hongguo, and Yanhui (2018) | | x | | | | | | | x | |
| Yang and Cao (2019) | x | x | | | | | | | | |
| Puspasari et al. (2021) | x | | | | | | | | | |
| Kumar et al. (2020) | | | | | x | | | | | |
| Jiang (2019) | | | | | | | | | | |
| Hooshyar et al. (2020) | x | x | x | | x | x | | | | |
| Awujoola (2021) | x | x | | | | | | | | |
| Nie et al. (2017) | | x | | | | | | | | |
| Heriyadi (2021) | | | | | | | | | | x |
| Gunarathne et al. (2021) | | | | | | | | | | x |
| Kahn et al. (2019) | | | | | | | | | | x |
| Belsare and Deshmukh (2018) | | | | | | | | x | x | |
| Martinez-Gil et al. (2018) | | x | | | | | | | | |
| Sharma et al. (2021) | | x | | | | | | | | |
| Van Huynh et al. (2020) | | | | | x | x | | | | |
| Vassef et al. (2022) | x | x | | | x | | | | | |
| Zhu et al. (2021) | | | | | x | | | | | |
| Uddin (2022) | | | | | | | | | x | |
| Li, Chuancheng, Yinggang, et al. (2018) | | | | | | | | | x | |

**Table 2.4: Proposed methods by researchers (Part 2)**

| Author | Clustering | Classification | Regression | Association | Deep Learning | Ensemble Learning | Dimensionality Reduction | Filtering | Correlation/Similarity Measures | Statistical Data Analysis |
|---|---|---|---|---|---|---|---|---|---|---|
| Vinutha and Yogisha (2021) | | x | | | x | x | | | | |
| Bannaka et al. (2021) | | x | | | | x | | | | |
| Saeed et al. (2021) | | | | | x | | | | | |
| Almutairi and Hasanat (2018) | | x | | | | | | | | |
| Angesti et al. (2021) | | x | x | | | | | | | |
| Jamal et al. (2020) | | x | | | | | | | | |
| Nachev and Teodosiev (2018) | | x | | | | | | | | |
| Amalia1 and Wibowo (2020) | | x | | | | | | | | |
| Arafath et al. (2018) | | x | x | | x | | | | | |
| N. VidyaShreeram1 (2021) | | x | | | | x | | | | |
| Ankush Daharwal1 (2020) | | x | | | | | | | | |
| Kamal et al. (2021) | | x | | | | x | | | | |
| Bharambe et al. (2017) | | x | | | | x | | | | x |
| Sui et al. (2018) | | | | | | | | | | x |
| D.Haritha (2019) | | x | | | | x | | | | |
| Qin et al. (2018) | | x | | | x | x | | | | |

Table 2.3 and 2.4 shows the proposed solution as supposed to the challenges that is stated on section 2.1, domain problems tackled by researchers. From the table above, we can see that the most proposed solution is classification which is 26 out of 50. Besides, the second most proposed solution is Deep Learning which proposed by 11 out of 50 researchers. In addition, ensemble learning and clustering were ranked in third place as 10 out of 50 papers and fourth place as 9 out of 50 papers respectively for the most proposed method. Meanwhile, the least proposed solution as only 1 out of 50 papers had covered the topic of workplace preference analytics, least of the researchers are particularly focused on Association and Dimension reduction.

## 2.4 Proposed techniques by researchers

Following a review of the proposed method, there are several techniques that correspond to each method. Nonetheless, in order to achieve a better outcome, each technique was chosen based on the challenges targeted, and the model's reliability and applicability. Regarding the most chosen technique, the solution would be sound to be implemented through this project.

**Table 2.5: Techniques of Classification**

| Author | Random Forest | Decision Tree | Naive Bayes | Support Vector Machine | K-Nearest Neighbor | Logistic Regression | Linear Discriminant Analysis |
|---|---|---|---|---|---|---|---|
| Fan (2020) | | x | | | | | |
| Yu and Zhang (2018) | | x | | | | | |
| Premalatha and Sujatha (2021) | | x | x | | | | |
| Nudin et al. (2022) | x | | | | | | |
| Chakraborty et al. (2019) | | x | | | | | |
| Yang and Cao (2019) | | | | | | | x |
| Hooshyar et al. (2020) | x | x | x | x | | | |
| Awujoola (2021) | x | x | x | x | | x | |
| Nie et al. (2017) | x | x | | x | | x | |
| Martinez-Gil et al. (2018) | x | | | x | | | |
| Sharma et al. (2021) | x | | | | | | |
| Vassef et al. (2022) | | | | | | x | |
| Vinutha and Yogisha (2021) | x | | x | x | x | x | |
| Bannaka et al. (2021) | x | x | x | x | x | | |
| Almutairi and Hasanat (2018) | | x | x | | x | | |
| Angesti et al. (2021) | | x | | | | | |
| Jamal et al. (2020) | | x | x | | | | |
| Nachev and Teodosiev (2018) | | | | x | | | |
| Amalia1 and Wibowo (2020) | | | x | | | | |
| Arafath et al. (2018) | x | x | | x | | | |
| N. VidyaShreeram1 (2021) | x | x | | x | | | |
| Ankush Daharwal1 (2020) | | | x | | | | |
| Kamal et al. (2021) | x | | | | | | |
| Bharambe et al. (2017) | x | x | | | x | | |
| D.Haritha (2019) | x | x | x | x | x | x | x |
| Qin et al. (2018) | x | x | | | | x | |

Table 2.5 shows the proposed techniques as supposed to the classification method that is stated in Tables 2.3 and 2.4. According to the table above, the Decision Tree is the most frequently proposed classification technique, accounting for 15 out of the 26 papers, more than half of the total. Moreover, the second most proposed is Random Forest which accounts for 14 out of 26 papers.

Naive Bayes and Support Vector Machines are the two techniques that came in third place in terms of most proposed techniques, with 10 out of 26 papers each. Meanwhile, the least proposed technique is Linear Discriminant Analysis (LDA) as only 2 out of 26 papers.

**Table 2.6: Techniques of Deep Learning**

| Author | Neural Network (NN)/ Artificial Neural Network (ANN) | Recurrent Neural Network (RNN) | Convolutional neural network (CNN) | TextCNN | Multilayer perceptron (MLP) | Long Short-Term Memory (LSTM) | BLSTM-A | PLSTM | Deep Neural Network (DNN) | Bi-GRU-LSTM-CNN / Bi-GRU-CNN |
|---|---|---|---|---|---|---|---|---|---|---|
| Nigam et al. (2019) | | | | | | | x | | | |
| Premalatha and Sujatha (2021) | | | | | x | | | | | |
| Kumar et al. (2020) | | x | | | | x | | | | |
| Hooshyar et al. (2020) | x | | | | | | | | | |
| Van Huynh et al. (2020) | | | | x | | | | | | x |
| Vassef et al. (2022) | | | x | | | | | | x | |
| Zhu et al. (2021) | | | | | | x | | x | | |
| Vinutha and Yogisha (2021) | x | | | | | | | | | |
| Saeed et al. (2021) | x | | x | | | | | | | |
| Arafath et al. (2018) | | | | | x | | | | | |
| Qin et al. (2018) | | x | | | | | | | | |

Table 2.6 shows the proposed techniques in relation to the deep learning method stated in Tables 2.3 and 2.4. Based on the table above, the Neural Network (NN)/Artificial Neural Network (ANN) is the most proposed deep learning technique, appearing in 3 out of the 11 papers. Besides that, the second most frequently proposed techniques are Recurrent Neural Network (RNN), Convolutional neural network (CNN), Multilayer perceptron (MLP) and Long Short-Term Memory (LSTM), which accounts for 2 out of 11 papers. In the meantime, the rest techniques only have 1 out of 11 papers to be proposed by researchers.

**Table 2.7: Techniques of Ensemble Learning**

| Author | AdaBoost (ADB) | Xgboost | Gradient Boosting | Stacking | Bagging |
|---|---|---|---|---|---|
| Premalatha and Sujatha (2021) | | | | x | x |
| Hooshyar et al. (2020) | x | | | | |
| Van Huynh et al. (2020) | | | | x | |
| Vinutha and Yogisha (2021) | | x | x | | |
| Bannaka et al. (2021) | | x | | | |
| N. VidyaShreeram1 (2021) | x | | | | |
| Kamal et al. (2021) | | x | | | |
| Bharambe et al. (2017) | x | | | | |
| D.Haritha (2019) | x | | | | |
| Qin et al. (2018) | x | | x | | |

Table 2.7 summarize the proposed techniques in accordance with the ensemble learning method listed in Tables 2.3 and 2.4. As shown in the table above, the boosting method was most frequently proposed in ensemble learning, with AdaBoost (ADB) appearing in 5 of the 10 papers and Xgboost appearing in 4 of the 10 papers. Bagging is the least proposed technique as only 1 out of 10 papers.

**Table 2.8: Techniques of Clustering**

| Author | K-means | K.DBSCAN | Probabilistic Clustering and Prediction (PCP) |
|---|---|---|---|
| Chen et al. (2019) | x | | |
| Rahman and Asadujjaman (2021) | x | | |
| Megasari et al. (2020) | x | | |
| Zhou et al. (2019) | x | | |
| Yang and Cao (2019) | | x | |
| Puspasari et al. (2021) | x | | |
| Jiang (2019) | | | x |
| Hooshyar et al. (2020) | x | | |
| Vassef et al. (2022) | x | | |

Table 2.8 discuss the proposed techniques as supposed to the clustering method stated in Tables 2.3 and 2.4. Looking at the table above, k-means was proposed as the most frequent technique which shows up in 7 out of 9 papers. In contrast, K.DBSCAN and Probabilistic Clustering and Prediction (PCP) are the least proposed technique as only 1 out of 9 papers.

Aside from the techniques listed in the table above, the rest techniques were proposed by the least of the papers. In the method of correlation/similarity measures, only similarity measure was proposed by 3 out of 8 papers. Only a small number of studies presented the remaining methods used in this method including Fuzzy Sugeno, Pearson correlation, Jaccard similarity coefficient, Cosine similarity, Vector Space Model (VSM) and feature selection/extraction which were only proposed in 1 or 2 papers of each technique.

Besides that, statistical data analysis has also been suggested for matching the criteria of graduates and the workplace that suit them. Among 6 papers, 1 or 2 pa-

pers each proposed the use of Quadratic Discriminant Analysis (QDA), Descriptive quantitative analysis, Bibliometric methods and Statistical Analysis.

Moreover, a small portion of researchers has recommended the regression method as well. In this method, only CART (Classification and Regression Tree) was proposed by 2 out of 4 papers, the remaining techniques such as Simple Logistics (SL) and Gaussian processes (GP) only suggested by 1 paper each.

Lastly, the rest 3 techniques, Collaborative filtering (CF), PCA (Principal Component Analysis) and Association rules mining were advocated by the fewest papers. Only Collaborative filtering (CF) was recommended twice in 4 papers but the rest were each recommended once.

To sum up, the summary of this chapter should focus on taking the challenges, methods and techniques that used by the majority of researchers as a reference for constructing a good predictive model in this project. In a nutshell, the research solution has the potential to serve as a guide and key finding that improves project preparation and process in advance. In contrast, the least case proposed by the researchers should also be taken into consideration if necessary based on special situations but in low priority.

# CHAPTER 3

# METHODOLOGY

## 3.1 Introduction

Workplace preference analytics refers to the use of data and analysis to understand the preferences and priorities of graduates when it comes to their ideal workplaces. This can include factors such as the location of the company, the sector of work and so on. Analyzing this data can help to understand what attracts graduates to certain companies and what they value in a workplace, which can inform the selection strategies for their career path. In order to meet the requirement of this project, a clear approach and strategy should be thoroughly outlined. In this chapter, the flow and techniques of constructing the entire system will be explained in depth for each stage.

Before diving into the in-depth description of each stage, a quick overview of the project framework and flow will be given in this section. We will be starting by collecting raw data, and then performing data preparation and data preprocessing to improve the data quality. After that, exploratory data analysis will be carried out for discovering data patterns and associations between the variables such as data visualization. Before the stage of building a predictive model, feature selection is an essential step to rank the optimal feature in the dataset given. Followed by performing data splitting, therefore we will have training data to perform model construction by training the model and also have testing data to perform model evaluation by inputting unseen data. After getting the result, the chosen predictive model will then be deployed and integrated into an application to be ready for production, as following the flowchart of the project framework construction below:

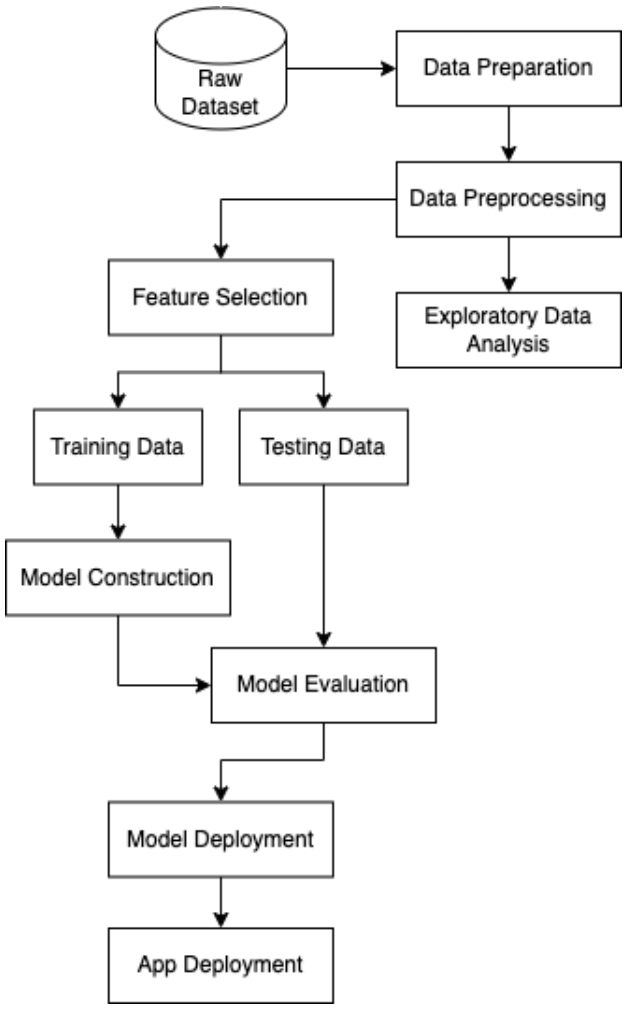


**Figure 3.1: Flowchart of Framework Construction**

## 3.2 Data Preparation

Since the current dataset is sourced by a university as real-world data, the data collection is completed at this stage, thus understanding the structure and definition of the data are essential for better analysis and optimization later on. It's important to note that data gathering, merging, structuring and organizing are all included in this process. The analysis work of this project is supplemented by analyzing two extra datasets from other sources.

### 3.2.1 Graduate Dataset

For this research, a graduate dataset is absolutely essential. Thus, our primary dataset is provided by a Malaysian private university. The dataset consists of a profile of 1754 graduates with 251 columns which represent their personal information including course name, graduation date, results of each trimester, score of CGPA (Culmulative Grade Point Average), current job status, present workplace location and much more. These graduates completed their respective programmes on dates between 7 April 2019 and 1 August 2021.

| | NewID | ACAD_CAREER | PROG_STATUS | PROG_ACTION | STATUS_DT | ADMIT_TERM | BEGIN_DT | END_DT | EXP_GRAD_TERM | CAMPUS | ... | daerah_sya |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | G1010910 | UGRD | CM | COMP | Thursday, 6 May, 2021 | 1730 | Monday, 9 April, 2018 | Sunday, 11 April, 2021 | 2020 | MLAKA | ... | no |
| **1** | G1015423 | UGRD | CM | COMP | Thursday, 6 May, 2021 | 1730 | Monday, 9 April, 2018 | Sunday, 11 April, 2021 | 2020 | MLAKA | ... | no |
| **2** | G1008210 | UGRD | CM | COMP | Monday, 24 May, 2021 | 1720 | Monday, 20 November, 2017 | Sunday, 5 July, 2020 | 1930 | MLAKA | ... | no |
| **3** | G1004436 | UGRD | CM | COMP | Thursday, 2 September, 2021 | 1730 | Monday, 9 April, 2018 | Sunday, 11 April, 2021 | 2020 | MLAKA | ... | no |
| **4** | G1006594 | UGRD | CM | COMP | Wednesday, 13 January, 2021 | 1710 | Monday, 3 July, 2017 | Sunday, 22 November, 2020 | 2010 | MLAKA | ... | no |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **1749** | G1007340 | UGRD | CM | COMP | Thursday, 12 August, 2021 | 1630 | Monday, 3 April, 2017 | Sunday, 1 August, 2021 | 2030 | CYBER | ... | Tiada Da K.Lu |
| **1750** | G1012209 | UGRD | CM | COMP | Thursday, 12 August, 2021 | 1630 | Monday, 3 April, 2017 | Sunday, 1 August, 2021 | 2030 | CYBER | ... | Daerah Pe |
| **1751** | G1003687 | UGRD | CM | COMP | Thursday, 12 August, 2021 | 1630 | Monday, 3 April, 2017 | Sunday, 1 August, 2021 | 2030 | MLAKA | ... | Daerah M Te |
| **1752** | G1001799 | UGRD | CM | COMP | Thursday, 12 August, 2021 | 1630 | Monday, 3 April, 2017 | Sunday, 1 August, 2021 | 2030 | MLAKA | ... | Daerah M Te |
| **1753** | G1002826 | UGRD | CM | COMP | Thursday, 8 April, 2021 | 1820 | Monday, 19 November, 2018 | Sunday, 11 April, 2021 | 2020 | CYBER | ... | Daerah Pe |

1754 rows × 251 columns

**Figure 3.2: Graduate Dataset**

### 3.2.2 Graduate Location Analytical Dataset

A geographically-focused analytical dataset can be built upon the existing base dataset based on the location of the graduates. In order to obtain the feature properties surrounding the graduates' living place, relative wealth index and population of nearby properties, a number of external datasets are needed to perform the analytics. However, this would be extremely beneficial for the next step of the predictive model. The new analytical dataset eventually contains 1748 rows of graduates' data with 567 columns after merging with the geographic data. There is a need to take note that the reason causing the total number of graduates to be reduced in this dataset is the removal of duplicated graduates in the phase of data cleaning.

| | lat | lng | KFC | 7 Eleven Malaysia Sdn Bhd | Pizza Hut | McDonalds | Dominos | 99 Speed Mart | 99 Speedmart Sdn Bhd | Dommal Food Services Sdn Bhd | MarryBrown | Starbucks | Surau | Restaurants | Proper Manageme |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2.275412 | 102.216042 | 1.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 8.0 | 5 |
| **1** | 2.152746 | 102.569608 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 7.0 | 5 |
| **2** | 2.242628 | 102.214621 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 6.0 | 1 |
| **3** | 3.084866 | 101.740811 | 1.0 | 1.0 | 1.0 | 0.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.0 | 1.0 | 0.0 | 101.0 | 86 |
| **4** | 2.425797 | 102.677138 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 3.0 | 1 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **1743** | 2.929784 | 101.703427 | 1.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 21.0 | 10 |
| **1744** | 3.079804 | 101.481770 | 1.0 | 1.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 32.0 | 23 |
| **1745** | 2.236954 | 102.213203 | 1.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 3.0 | 4 |
| **1746** | 2.256872 | 102.252533 | 1.0 | 0.0 | 1.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 9.0 | 6 |
| **1747** | 3.192620 | 101.775337 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 8.0 | 15 |

**Figure 3.3: Graduate Location Analytical Dataset**

### 3.2.3 Company Location Analytical Dataset

With the presence of base dataset, an analytical dataset can be constructed based on the company location as well. This analytic acquires the nearby properties of the current workplace of the graduates, relative wealth index and population of nearby properties surrounding the company which the concepts are similar to the graduate analytical dataset in Section 3.2.2. After combining the company data with the geographic data, a new company location analytical dataset finally been constructed which contains 1748 rows of company's data with 567 columns based on graduates per row.

| | lat | lng | KFC | 7 Eleven Malaysia Sdn Bhd | Pizza Hut | McDonalds | Dominos | Starbucks | Nandos | Kenny Rogers Roasters | Mynews Retail Sdn Bhd | Dommal Food Services Sdn Bhd | 99 Speedmart Sdn Bhd | Restaurants | Consultant |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| **1** | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| **2** | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| **3** | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| **4** | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **1743** | 3.118320 | 101.677120 | 1.0 | 1.0 | 0.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 198.0 | 210. |
| **1744** | 3.086024 | 101.588821 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 116.0 | 130. |
| **1745** | 2.235587 | 102.259576 | 1.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 10.0 | 5. |
| **1746** | 2.227913 | 102.226439 | 1.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 10.0 | 4. |
| **1747** | 3.104103 | 101.642800 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 132.0 | 145. |

**Figure 3.4: Company Location Analytical Dataset**

## 3.3 Data Preprocessing

After getting an adequate quantity of data from a particular target audience, thus allowing data preprocessing begins with the aim of enhancing data quality, consistency, and completeness. It is the procedure of manipulating raw data into a much-desired form that makes it ready to be precisely analysed. This is an essential step in data mining to minimize the impact of the case like manual input errors, missing data, repetitive data and so on, to ensure this process resulting better performance of the predictive model.

### 3.3.1 Data Cleaning

Data cleaning is an extremely essential early step in the data analytics process which defines the quality of the dataset. A tedious task to ensure the completeness and accuracy of the data by fixing unreasonable data and error detection by cleaning up the dirty data. First of all, duplicated data often occur during data collection which is a vital problem that should get rid of them at the first step. It is because duplicated data is harmful to the accuracy of data such as conflicting the exact count of each record represented within a dataset. Therefore, the removal of duplicated data was executed on this dataset.

The second issue of the dataset is dealing with missing data. Different terms like "no_data" and "Irrelevant" were found in the dataset which has the same definition as missing values or even remaining blank cells. In this case, the solution needs to be selected depending on the situation. For instance, dropping the rows containing a missing value on featured columns like missing student ID might assume the student does not exist. Besides that, filling in some missing values with the default value manually can be executed as well. For example, the missing data that are in numerical form was filled in with the mean of that particular column such as Age Range. Additionally, imputing the data by using a function named SimpleImputer with the strategy called most frequent, which replaces missing values with the most frequent value along each

column. As a consequence, it is possible to avoid any potential bias in the results.

| skop_pekerjaan | bekerja_dalam_bidang_sama_belajar | tempoh_menunggu_pekerjaan_pertama | sebab_belum_bekerja | syarikat_latitude | syarikat_longitude |
|---|---|---|---|---|---|
| no_data | no_data | no_data | Awaiting job placement (have received a job of... | 0.000000 | 0.000000 |
| no_data | no_data | no_data | Looking for a job | 0.000000 | 0.000000 |
| no_data | no_data | no_data | no_data | 0.000000 | 0.000000 |
| no_data | no_data | no_data | no_data | 0.000000 | 0.000000 |
| no_data | no_data | no_data | no_data | 0.000000 | 0.000000 |

**Figure 3.5: Missing Values**

Thirdly, inconsistent data will be flawed for prediction later on. The inconsistent data shown in the figure below is hard to be categorized and converted through data transformation since it is a bunch of string characters with commas. In this case, 'INFO1' is actually made up of the SPM result of different subjects gotten by the graduates. In this way, each subject should be splitted into separate columns and assign the score into the specific column. In addition, as "Bahasa Cina(SPM)" and "Bahasa Cina(SPM-2013)" exist which signify the same thing, therefore the data of both columns were merged and the year 2013 was removed since it adds nothing of significance and it is not a key value for the dataset.

| | INFO1 |
|---|---|
| **0** | Bahasa Cina(SPM-2013):B+, Bahasa Inggeris(SPM-... |
| **1** | Bahasa Cina(SPM-2013):B, Bahasa Inggeris(SPM-2... |
| **2** | Bahasa Inggeris(SPM):B, Bahasa Melayu(SPM):A, ... |
| **3** | MUET(MUET):3 |
| **4** | MUET(MUET):3 |
| **...** | ... |
| **1749** | MUET(MUET):4 |
| **1750** | MUET(MUET):4 |
| **1751** | MUET(MUET):4 |
| **1752** | Bahasa China / Chinese Language(UEC):C8, Bahas... |
| **1753** | Bahasa Inggeris(SPM):A, Bahasa Melayu(SPM):B+,... |

**Figure 3.6: Inconsistent Data**

Figure below shows the splitted subject into separated columns and the data of each row is not complete which needs further execution on handling missing data.

| | Bahasa Cina(SPM) | Ekonomi Asas(SPM) | Matematik(SPM) | Matematik Tambahan(SPM) | Prinsip Akaun(SPM) | Sains(SPM) | Sejarah(SPM) | Biologi(SPM) | Fizik(SPM) | Kimia(SPM) | Pengetahuan Moral(SPM) | Pe Se |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | B+ | A | A+ | A- | A+ | A+ | A | NaN | NaN | NaN | NaN | |
| 1 | B | NaN | A | A- | NaN | NaN | A- | B+ | B+ | A- | NaN | |
| 2 | NaN | NaN | D | G | NaN | NaN | A- | C | D | D | NaN | |
| 3 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | |
| 4 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| 1749 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | |
| 1750 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | |
| 1751 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | |
| 1752 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | |
| 1753 | NaN | NaN | A | C+ | C+ | NaN | B | C | D | D | NaN | |

**Figure 3.7: Splitted Inconsistent Data**

## 3.4 Model Construction

In this project, a cluster analysis and classification model will be developed for predicting the workplace preference for graduates. In this project, a cluster analysis and classification model will be developed for predicting the workplace preference for graduates. Since K-means clustering is the most widely advocated method for cluster analysis by researchers, thus it is chosen in this project. In contrast, several classification models are chosen based on how well regarded and ranked they were among researchers and how well the models fit the project's needs. We will implement K-Nearest Neighbor, Decision Tree, Random Forest, Naive Bayes and Support Vector Machine as the predictive models for this project.

## 3.5 Model Evaluation

Model evaluation is crucial to specifically assess the performance of a classification model after it has been built which involves measuring how well the model is able to make accurate predictions on unseen data. Evaluation method selection is ultimately determined by the nature of the problem at hand and the available data. It's always recommended to employ multiple methodologies to get a better picture of the model's performance from several perspectives. There are several ways to evaluate a single machine learning model which confusion matrix, ROC curve and metrics like accuracy, precision, recall, and F1 score.

To understand how the calculation of metrics and confusion matrix work on evaluation, there are 4 possible outcomes in the context of classification problems:

- True Positive (TP): An instance where the model correctly predicts the positive class. For example, the model correctly identifies a patient as having a disease.
- False Positive (FP): An instance where the model incorrectly predicts the positive class. For example, the model incorrectly identifies a healthy patient as having a disease.
- True Negative (TN): An instance where the model correctly predicts the negative class. For example, the model correctly identifies a healthy patient as not having a disease.
- False Negative (FN): An instance where the model incorrectly predicts the negative class. For example, the model incorrectly identifies a patient as not having a disease when the patient actually does have the disease.

### 3.5.1 Accuracy

Accuracy is one metric to evaluate a classification model's performance. Accuracy can be thought of as the percentage of correct predictions for the test data. It is a simple calculation by dividing the number of correct predictions by the number of total predictions as the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

**Figure 3.8: Formula of Accuracy Calculation**

Although we assess the performance of the model mostly depending on the accuracy of the test set which is for the unseen data, it is best to verify the accuracy on the training set as well when making comparisons of multiple models to get a whole picture. It is theoretically possible to get a 1.0 value for accuracy on the training set (previously observed data) which scores 100% or so-called perfect accuracy, but it is generally not a good indicator of the model's performance. This is because the model has already seen the training data and has been optimized to fit that data perfectly. In contrast, it is not achievable to get 1.0 on the test set which consists of data that has not been seen by the model. Overfitting may lead to unrealistically high accuracy hence a model checking should be conducted in this scenario.

### 3.5.2 Precision

Precision as known as positive predictive value is the next place to be inspected in metric. Precision is the proportion of correctly predicted positive classifications from the cases that are predicted to be positive. In other words, precision measures how often a model predicts correctly in the positive class as the following formula:

$$Precision = \frac{TP}{TP + FP}$$

**Figure 3.9: Formula of Precision Calculation**

A great classifier should have a precision value that is close to 1 which is considered high. If the numerator and denominator are equal as TP = TP +FP shown in the formula, then FP equals 0, hence the precision value will become 1 which is the perfection. As a consequence, the increase of FP will cause the value of the denominator greater than the numerator which resulting the precision value decreasing and leading to worse performance.

### 3.5.3 Recall

Recall is also known as sensitivity or true positive rate. Recall is the proportion of correctly predicted positive classifications from the cases that are actual positives. In other words, recall measures how frequently a model predicts a positive when it is truly positive as the following formula:

$$Recall = \frac{TP}{TP + FN}$$

**Figure 3.10: Formula of Recall Calculation**

Similar concept with precision, a recall value that is close to 1 indicates a good classifier. If the numerator and denominator are identical as TP = TP +FN shown in the formula, then FN equals 0, hence the precision value ultimately scores 1 which is the highest value. Consequently, when FP increases, the denominator value becomes larger relative to the numerator which causes the recall value to drop and negatively impacts the performance.

### 3.5.4 F1-Score

F1 score is a weighted average of precision and recall. Keeping an eye on the F1 score could be the optimal way to strike the best possible balance between recall and precision. In most cases, the F1 score is more beneficial than accuracy, particularly when dealing with uneven class distribution. Accuracy is useful when there are similar costs in FP and FN, but it's suggested to pay close attention to both precision and recall the difference between the cost of FP and FN are vary greatly. The formula used for F1-score is as following:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

**Figure 3.11: Formula of F1-Score Calculation**

By taking into account both precision and recall, when a classifier score a perfect F1 score when the value is 1, it indicates precision and recall are both 1 too, the classifier is said to be good. Hence, the F1 score is highly dependent on the ratio of precision and recall to rise. However, there is a limitation of the F1 score which does not take into consideration the TN like how accuracy work.

### 3.5.5 Confusion Matrix

Confusion matrix is a matrix used to identify the classifier performance by visualizing the count of four values (TP, FP, TN, FN). The visual is mainly used to compare the count of actual class instances against predicted class instances. This allows the examination of the amount of correct and incorrect predictions for each category and the presence of any bias, and where it occurs if it exists. A confusion matrix is formed by plotting them against each other:

TP and TN are to be said the most concentrated part in the confusion matrix

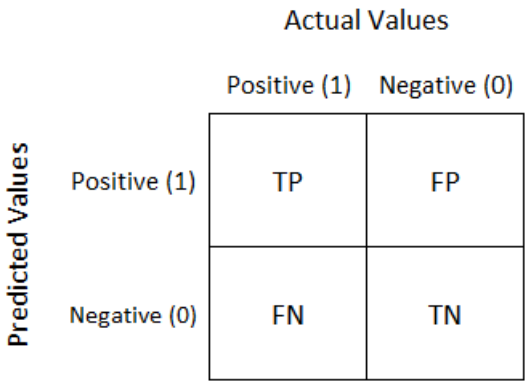


**Figure 3.12: Visual of Confusion Matrix**

which shows how well a model predicted the outcome correctly. On the other hand, FP as known as type I error and FN as known as type II error; both are the spots that show non-ideal results where the incorrect outcome falls on. It indicates how poor the performance of the model based on the count of the predicted value.

### 3.5.6 ROC Curve

Receiver operating characteristic (ROC) curve illustrates the trade-off between sensitivity and specificity which is commonly plotted to compare the area of various models under the curve (AUC). ROC curve typically is plotted based on the true positive rate (TPR) on the y-axis and false positive rate (FPR) on the x-axis which calculated as the following formula:

$$\mathrm{TPR} \stackrel{\mathrm{def}}{=} \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \text{ and } \mathrm{FPR} \stackrel{\mathrm{def}}{=} \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}.$$

**Figure 3.13: Formula of ROC Curve Calculation**

The AUC is the area under the ROC curve. Hence, the larger the area under the ROC curve, the closer the AUC value is to 1 and the greater the performance of the

model is. In other words, a classifier that provides curves closer to the top-left corner indicates a better performance.

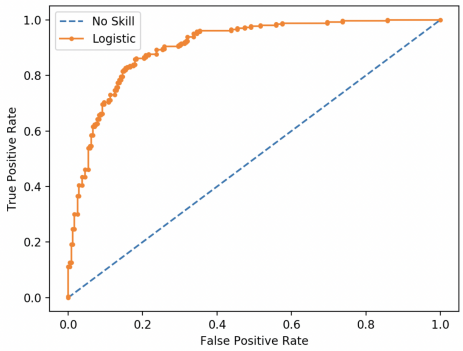


**Figure 3.14: ROC Curve Plot**

As the ideal plot should arc close to the top-left corner of the chart, a good classifier should achieve an AUC greater than 0.5 which is better than a random classifier. If not, the model would have flaws in such case. An AUC of 1 would indicate a perfectly accurate classifier. Referring to the figure of the ROC curve plot above, the logistic regression is a good classifier which nearly reached an AUC of 0.8 in this case. Lastly, AUC provides a decent overall overview of the classifier's predictive ability, when the dataset is imbalanced.

# CHAPTER 4

# FINDINGS

## 4.1 Introduction

Data visualization is a crucial part which allows a data analyst to dive into pattern discovery and a better understanding of the relationship between the variables. In this project, Microsoft Power BI was used to create an interactive dashboard with the data visualization approach to explore and analyse information from visuals. Power BI is a business intelligence platform commonly used to find insights within data of an organization such as creating charts and graphs. There are 4 personalized dashboards created and hosted on the Power BI server which allows users to access them in a quick and simple way by just clicking the provided URL.

## 4.2 Data Visualization

Data visualization is an approach that most businesses looking for to analyze data and share useful insight, in order to achieve the goal of making data easier to understand and more accessible. Data visualization commonly is to be said a graphical representation of data and information with visual elements like charts, graphs, and maps. It's a great tool for businesses to present their data to non-technical audiences in the least confusing way.

A graduate dashboard was built based on the information on their profile. This dashboard mainly indicates the geographical distribution of graduates on a map, what faculty they studied under, the distribution of gender, their qualifications when they first enter the institution and the field of study pursued by each gender.

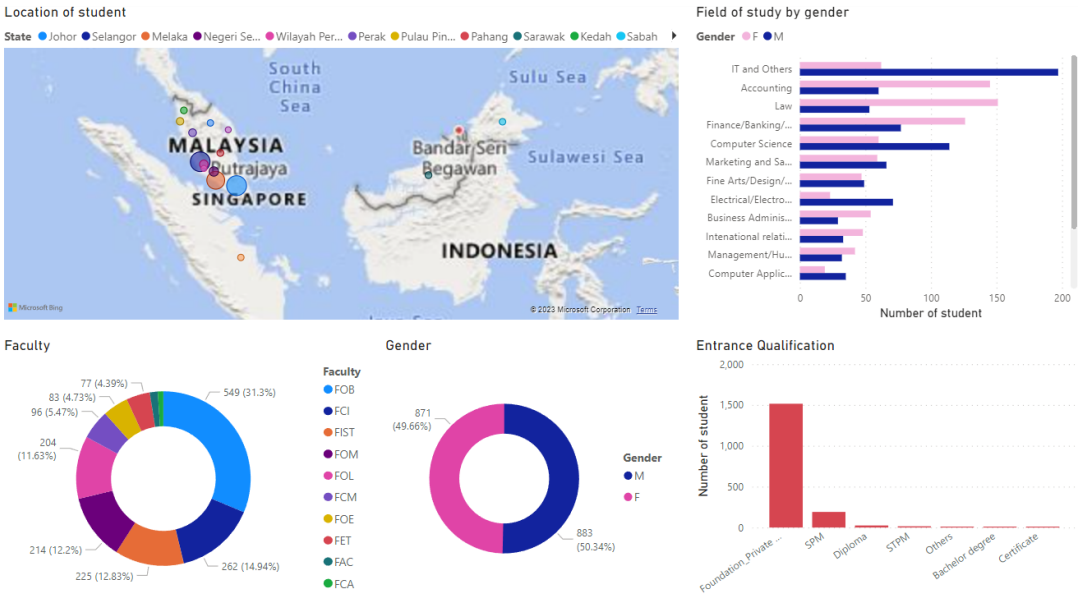


**Figure 4.1: Graduate Dashboard**

A graduate location dashboard was created and the information is taken from the graduate location analytical dataset. It presents the distribution of the top 10 featured buildings and the sum of feature nearby properties surrounding the living place of graduates. By clicking a spot of the location on the map, the total number of the top 10 featured buildings located in each spot will be shown.

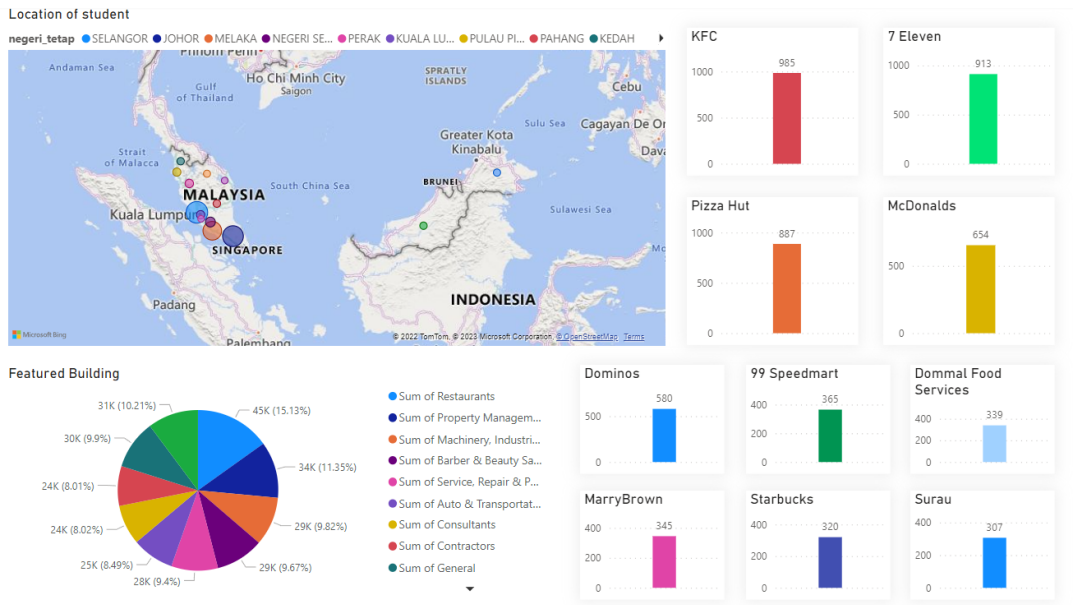


**Figure 4.2: Graduate Location Dashboard**

A company dashboard was presented with the distribution of the company located geographically, the industry sector each gender had joined upon graduation, the employment status of each graduate as well as cgpa of each graduate to find a job.

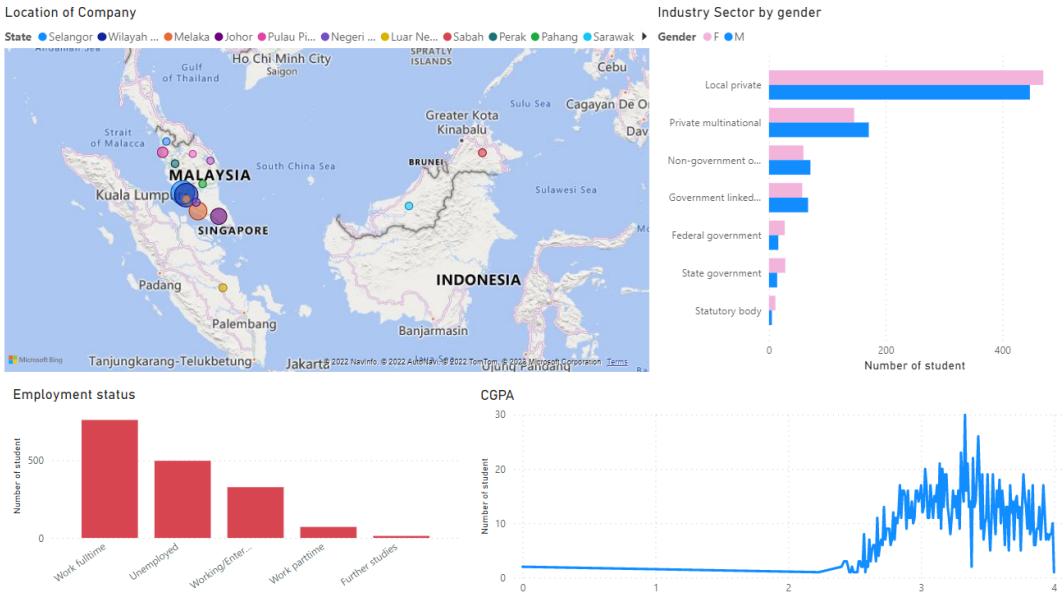


**Figure 4.3: Company Dashboard**

Lastly, a company location dashboard was constructed for visualizing the company location analytical dataset. There is a map showing the distribution of company locations, along with the count of the top 10 featured buildings surrounding it and the total nearby properties surrounded such as restaurants, consultants and so on.

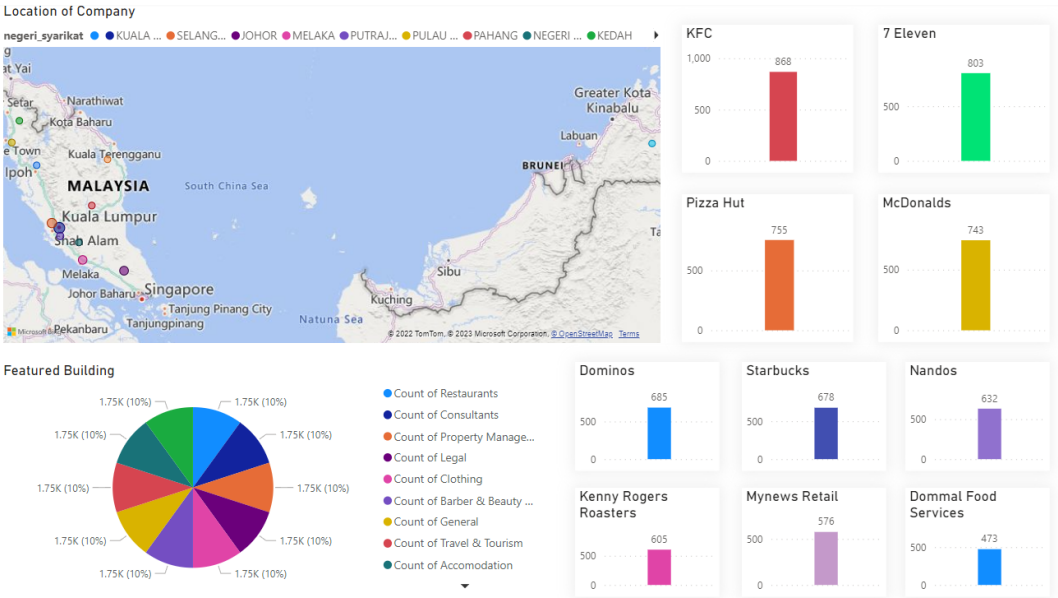


**Figure 4.4: Company Location Dashboard**

# CHAPTER 5

# IMPLEMENTATION PLAN

## 5.1 Gantt Chart

The figures below shows the Gantt Chart of FYP 1 and FYP 2.

Gantt Chart for FYP 1 (Trimester 1, 2022/2023)

| TASKS | WEEK | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| Initiation | | | | | | | | | | | | | | | | |
| Project Planning | | | | | | | | | | | | | | | | |
| Literature Review Study | | | | | | | | | | | | | | | | |
| Dataset | | | | | | | | | | | | | | | | |
| Primary Dataset - Data Preparation | | | | | | | | | | | | | | | | |
| Secondary dataset - Data Preparation | | | | | | | | | | | | | | | | |
| Data preprocessing | | | | | | | | | | | | | | | | |
| Power BI - Data Visualization | | | | | | | | | | | | | | | | |
| Report Writing | | | | | | | | | | | | | | | | |
| Chapter 1: Introduction | | | | | | | | | | | | | | | | |
| Chapter 2: Literature Review | | | | | | | | | | | | | | | | |
| Chapter 3: Methodology | | | | | | | | | | | | | | | | |
| Chapter 4: Findings | | | | | | | | | | | | | | | | |
| Chapter 5: Implementation Plan | | | | | | | | | | | | | | | | |
| Chapter 6: Conclusion | | | | | | | | | | | | | | | | |
| Report Review by Supervisor | | | | | | | | | | | | | | | | |
| Interim Report Submission | | | | | | | | | | | | | | | | |
| FYP1 Presentation | | | | | | | | | | | | | | | | |

**Figure 5.1: Gantt Chart of FYP1**

Gantt Chart for FYP 2 (Trimester 2, 2022/2023)

| TASKS | WEEK | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| Implementation | | | | | | | | | | | | | | | | |
| Exploratory Data Analysis | | | | | | | | | | | | | | | | |
| Feature Selection | | | | | | | | | | | | | | | | |
| Model Construction | | | | | | | | | | | | | | | | |
| Model Evaluation | | | | | | | | | | | | | | | | |
| Model & App Deployment | | | | | | | | | | | | | | | | |
| Writing | | | | | | | | | | | | | | | | |
| Research Paper | | | | | | | | | | | | | | | | |
| FYP2 Report | | | | | | | | | | | | | | | | |
| Research Paper Review by Supervisor | | | | | | | | | | | | | | | | |
| Report Review by Supervisor | | | | | | | | | | | | | | | | |
| FYP2 Report Submission | | | | | | | | | | | | | | | | |
| Oral Presentation | | | | | | | | | | | | | | | | |

**Figure 5.2: Gantt Chart of FYP2**

## 5.2 Project Target

Since a series of planning, preparation and research was done in phase of Final Year Project (FYP) 1, it will smoothen the stage of implementation on FYP 2. Based on the Gantt Chart of FYP 1, about a quarter completion of the entire project was reached with method study, data preparation and some minor data analysis. It is time to carry on the proposed flow of the framework as outlined in Section 3.1. Data preparation, data preprocessing and part of explanatory data analysis were done in FYP 1, next stage should be hands-on by constructing the predictive model with more explanatory data analysis, feature selection, evaluation, testing and deployment in FYP 2.

## 5.3 Additional Data Preprocessing

While working on more findings in the stage of data visualization as discussed in Chapter 4. A few more issues were encountered and found in the primary and secondary datasets due to dirty data, one example is the invalid home address which contains some strange characters. Hence, further data preprocessing is needed for enhancing the data quality for better predictive results during implementation.

## 5.4 Model Construction

Since research and study on techniques were done in the phase of FYP 1 through literature review, several machine learning algorithms will be trained and tested in the next phase, the predictive models including K-means clustering, K-Nearest Neighbor, Decision Tree, Random Forest, Naive Bayes and Support Vector Machine which used to predict the preferred workplace for graduates. An evaluation and ranking will be carried out in order to determince the best performance model for this project.

## 5.5 Web Application

The last stage of the project is to make the predictive model ready for production so more users can use it. Consequently, the high-performing models will be chosen and integrated into an application. In this project, FlutterFlow will be used to create a web application for graduates to find their desired workplace by just simply filling in some necessary profile information in a browser. FlutterFlow is a visual builder with Firebase integration, API support and more functionalities which allow a developer to build a fully functional app in a much easier and faster way.

# CHAPTER 6

# CONCLUSION

A great deal of work was accomplished in the phase in FYP 1. To recap, a general idea of the topic has been well discovered and spelt out, thus problems faced can be spotted thoroughly no matter real-world problems or technical issues in this project. In this way, the project's initiation phase started with project research and planning to determine the objectives, project scope and problem statement. Consequently, 50 research papers have been studied, the literature review could be taken as a guide or reference for the project direction based on the proven effectiveness of the existing method and the results of experiments tested by researchers. Afterwards, the challenges, proposed method and techniques from 50 papers were summarized and written in this report. Based on the literature review, the flow of the project framework was outlined with the requirement needed in this project.

Next, data preparation was started once the primary dataset was obtained from the Malaysian private university which consists of 1754 graduates' profiles. After that, 2 secondary datasets were constructed which are the graduate location analytical dataset and the company location analytical dataset in order to get the data of featured nearby properties. Moreover, data preprocessing was performed for the 3 datasets mentioned above to improve the data quality. Apart from that, some strategies for evaluating the predictive model were studied and described in this report to get a general idea of testing the model performance in the later stage. Additionally, the first finding that has been tried on the main dataset was through data visualization. Microsoft Power Bi allows users to view the demographics of graduates based on location, gender, field of study and many more, meanwhile distribution of companies also can be seen in relation to the profile of graduates, nearby properties and much more. However, definitely more explanatory data analysis should be tested in the next phase. The final task of the phase of FYP 1 is writing this interim report which in-

cludes the introduction of the project topic, literature review, methodology, findings, implementation plan and this section, conclusion.

In terms of unfinished tasks, it will be the goal of the next phase, FYP 2 to wrap up all the loose ends. Firstly, further data preprocessing is needed as several dirty data were discovered while processing the data in the data visualization stage. The second work is the most significant task of the project which is model construction. A few machine learning models will be trained and evaluated which should achieve an accuracy of at least 70% as a good predictive model and definitely take into account other aspects as well like precision, recall, roc-curve and many more studied from section 3.5. As previously said, the step of assessing model performance will be carried out in order to select the best-performing model for deployment. Lastly, a web application will be built by integrated with the best predictive model to make it publicly accessible to achieve the final goal of this project.

During the phase of FYP 1, there are quite a few problems were encountered. The first issue was unclear with the direction of choosing the suitable method for this project as there are so many approaches and procedures proposed. Therefore, the solution was to study at least 50 papers to adequately compare the proposed methods by the researchers. Secondly, when I was uncertain about how to extract the feature properties of a building based on an address. My supervisor helped me with some code as a reference and explanation, and then encourage me to form a study group with my classmate to work on it together. The third challenge was the interruption of dirty data while attempting to connect the dataset to the Power BI, hence necessitating more data processing was carried out to resolve this issue although it required extra effort and time-consuming to do the checking, which is essential for maintaining data quality.

# REFERENCES


[1] Almutairi, M. M., & Hasanat, M. H. A. (2018). Predicting the suitability of is students' skills for the recruitment in saudi arabian industry. In *2018 21st saudi computer society national computer conference (ncc)* (p. 1-6). doi: 10.1109/NCG.2018.8593016

[2] Amalia1, R., & Wibowo, A. (2020). Prediction of the waiting time period for getting a job using the naive bayes algorithm. In *International research journal of advanced engineering and science* (Vol. 5, p. 225-229).

[3] Angesti, R. G., Kurniawati, A., & Anggana, H. D. (2021). Prediction of the telkom university's undergraduates waiting period for getting a job using the cart algorithm. In *2021 4th international conference of computer and informatics engineering (ic2ie)* (p. 135-140). doi: 10.1109/IC2IE53219.2021.9649290

[4] Ankush Daharwal1, A. B. S. D. S. C., Prof. Sandeep Gore. (2020). Career guidance system using machine learning for engineering students (cs/it). In *International research journal of engineering and technology (irjet)* (Vol. 7, p. 3417-3420).

[5] Arafath, M. Y., Saifuzzaman, M., Ahmed, S., & Hossain, S. A. (2018). Predicting career using data mining. In *2018 international conference on computing, power and communication technologies (gucon)* (p. 889-894). doi: 10.1109/GUCON.2018.8674995

[6] Atela, R., Othuon, L., & Agak, J. (2020, 05). Relationship between types of intelligence and career choice among undergraduate students of maseno university, kenya.

[7] Awujoola, P. O. O. M. E. I. . H. A., O. (2021). Performance evaluation of machine learning predictive analytical model for determining the job applicants employment status. , *6*, 67-79. Retrieved from `https://journal.unisza.edu.my/myjas/index.php/myjas/article/view/276` doi: 10.37231/myjas.2021.6.1.276

[8] Bannaka, B. M. D. E., Dhanasekara, D. M. H. S. G., Sheena, M. K., Karunasena, A., & Pemadasa, N. (2021). Machine learning approach for predicting career suitability, career progression and attrition of it graduates. In *2021 21st international conference on advances in ict for emerging regions (icter)* (p. 42-48). doi: 10.1109/ICter53630.2021.9774825

[9] Belsare, R., & Deshmukh, D. (2018, 06). Employment recommendation system using matching, collaborative filtering and content based recommendation. *International Journal of Computer Applications Technology and Research*, *7*, 215-220. doi: 10.7753/IJCATR0706.1003

[10] Bharambe, Y., Mored, N., Mulchandani, M., Shankarmani, R., & Shinde, S. G. (2017). Assessing employability of students using data mining techniques. In *2017 international conference on advances in computing, communications and informatics (icacci)* (p. 2110-2114). doi: 10.1109/ICACCI.2017.8126157

[11] Chakraborty, D., Hossain, M. S., & Arefin, M. S. (2019). Demand analysis of cse graduates of different universities in job markets. In *2019 international conference on electrical, computer and communication engineering (ecce)* (p. 1-6). doi: 10.1109/ECACE.2019.8679511

[12] Chen, Z., Liang, W., Gao, X., Zhou, Z., & Wu, M. (2019). Research on the accurate recommendation management system for employment of college graduates on hadoop. In *2019 5th international conference on big data and information analytics (bigdia)* (p. 19-22). doi: 10.1109/BigDIA.2019.8802855

[13] D.Haritha. (2019). Smart career guidance and recommendation system. In *International journal of engineering development and research* (Vol. 7, p. 633-638).

[14] Fan, H. (2020). A prediction model of college students' employment based on data mining. In *2020 13th international conference on intelligent computation technology and automation (icicta)* (p. 549-552). doi: 10.1109/ICICTA51737.2020.00121

[15] Gunarathne, N., Senaratne, S., & Herath, R. (2021). Addressing the expectation–performance gap of soft skills in management education: An integrated skill-development approach for accounting students. *The International Journal of Management Education*, *19*(3), 100564. Retrieved from https://www.sciencedirect.com/science/article/pii/S1472811721001130 doi: https://doi.org/10.1016/j.ijme.2021.100564

[16] Heriyadi, B. (2021, 04). Tracer study analysis for the reconstruction of the mining vocational curriculum in the era of industrial revolution 4.0. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*, 3013-3019. doi: 10.17762/turcomat.v12i3.1335

[17] Hooshyar, D., Pedaste, M., & Yang, Y. (2020). Mining educational data to predict students' performance through procrastination behavior. *Entropy*, *22*(1). Retrieved from https://www.mdpi.com/1099-4300/22/1/12 doi: 10.3390/e22010012

[18] Jamal, K., Kurniawan, R., Husti, I., Zailani, Nazri, M. Z. A., & Arifin, J. (2020). Predicting career decisions among graduates of tafseer and hadith. In *2020 2nd international conference on computer and information sciences (iccis)* (p. 1-4). doi: 10.1109/ICCIS49240.2020.9257663

[19] Jiang, F. Y. X. H. e. a., M. (2019). User click prediction for personalized job recommendation. In *World wide web 22* (p. 325-345). doi: 10.1007/s11280-018-0568-z

[20] Jie, L., Zheng, S., Qi, W., & Xiya, C. (2021). Analysis of employment status and countermeasures of biology graduates in local normal universities based on big data technology—take the graduates of guangxi normal university from 20l6 to 2020 as an example. In *2021 2nd international conference on artificial intelligence and education (icaie)* (p. 572-578). doi: 10.1109/ICAIE53562.2021.00127

[21] Kahn, M., Gamedze, T., & Oghenetega, J. (2019). Mobility of sub-saharan africa doctoral graduates from south african universities—a tracer study. *International Journal of Educational Development*, *68*, 9-14. Retrieved from https://www.sciencedirect.com/science/article/pii/S0738059318308940 doi: https://doi.org/10.1016/j.ijedudev.2019.04.006

[22] Kamal, A., Naushad, B., Rafiq, H., & Tahzeeb, S. (2021). Smart career guidance system. In *2021 4th international conference on computing information sciences (iccis)* (p. 1-7). doi: 10.1109/ICCIS54243.2021.9676408

[23] Kumar, R. S., Prakash, N., & Anbuchelian, S. (2020). Prediction of job openings in it sector using long short -term memory model. In *2020 fourth international conference on i-smac (iot in social, mobile, analytics and cloud) (i-smac)* (p. 945-953). doi: 10.1109/I-SMAC49090.2020.9243483

[24] Kusnawi, K., Ipmawati, J., & Kusumandaru, D. (2019). Decision support system employee recommendation using fuzzy sugeno method as a job search service. In *2019 international conference on information and communications technology (icoiact)* (p. 539-542). doi: 10.1109/ICOIACT46704.2019.8938452

[25] Li, S., Chuancheng, Y., Hongguo, W., & Yanhui, D. (2018). An employment recommendation algorithm based on historical information of college graduates. In *2018 9th international conference on information technology in medicine and education (itme)* (p. 708-711). doi: 10.1109/ITME.2018.00161

[26] Li, S., Chuancheng, Y., Yinggang, L., Hongguo, W., & Yanhui, D. (2018). information to intelligence(itoi): A prototype for employment prediction of graduates based on multidimensional data. In *2018 9th international conference on information technology in medicine and education (itme)* (p. 834-836). doi: 10.1109/ITME.2018.00187

[27] Martinez-Gil, J., Freudenthaler, B., & Natschläger, T. (2018, 03). Recommendation of job offers using random forests and support vector machines..

[28] Megasari, R., Piantari, E., & Nugraha, R. (2020). Graduates profile mapping based on job vacancy information clustering. In *2020 6th international conference on science in information technology (icsitech)* (p. 35-39). doi: 10.1109/ICSITech49800.2020.9392067

[29] Nachev, A., & Teodosiev, T. (2018). Analysis of employment data using support vector machines. In *International journal of applied engineering research issn* (Vol. 13, p. 13525-13535).

[30] Nie, M., Yang, L., Sun, J., Su, H., Xia, H., Lian, D., & Yan, K. (2017, 10). Advanced forecasting of career choices for college students based on campus big data. *Frontiers of Computer Science*, *12*. doi: 10.1007/s11704-017-6498-6

[31] Nigam, A., Roy, A., Singh, H., & Waila, H. (2019). Job recommendation through progression of job selection. In *2019 ieee 6th international conference on cloud computing and intelligence systems (ccis)* (p. 212-216). doi: 10.1109/CCIS48116.2019.9073723

[32] Nudin, S. R., Warsito, B., & Wibowo, A. (2022). Impact of soft skills competencies to predict graduates getting jobs using random forest algorithm. In *2022 1st international conference on information system information technology (icisit)* (p. 49-54). doi: 10.1109/ICISIT54091.2022.9872669

[33] N. VidyaShreeram1, D. A. M. (2021). Student career prediction using machine learning approaches. In *I3cac 2021*. doi: 10.4108/eai.7-6-2021.2308642

[34] Olowolayemo, A., Harun, K., & Mantoro, T. (2018). University based job recommender alumni system. In *2018 international conference on computing, engineering, and design (icced)* (p. 212-217). doi: 10.1109/ICCED.2018.00049

[35] Premalatha, N., & Sujatha, S. (2021). An effective ensemble model to predict employment status of graduates in higher educational institutions. In *2021 fourth international conference on electrical, computer and communication technologies (icecct)* (p. 1-4). doi: 10.1109/ICECCT52121.2021.9616952

[36] Puspasari, B. D., Damayanti, L. L., Pramono, A., & Darmawan, A. K. (2021). Implementation k-means clustering method in job recommendation system. In *2021 7th international conference on electrical, electronics and information engineering (iceeie)* (p. 1-6). doi: 10.1109/ICEEIE52663.2021.9616654

[37] Qin, C., Zhu, H., Xu, T., Zhu, C., Jiang, L., Chen, E., & Xiong, H. (2018). Enhancing person-job fit for talent recruitment: An ability-aware neural network approach. In *The 41st international*

*acm sigir conference on research  development in information retrieval* (p. 25–34). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3209978.3210025 doi: 10.1145/3209978.3210025

[38] Rahman, M., & Asadujjaman, M. (2021). Multi-criteria decision making for job selection. In *2021 international conference on decision aid sciences and application (dasa)* (p. 152-156). doi: 10.1109/DASA53625.2021.9682215

[39] Rodriguez, L. G., & Chavez, E. P. (2019). Feature selection for job matching application using profile matching model. In *2019 ieee 4th international conference on computer and communication systems (icccs)* (p. 263-266). doi: 10.1109/CCOMS.2019.8821682

[40] Saeed, T., Sufian, M., Ali, M., & Rehman, A. U. (2021). Convolutional neural network based career recommender system for pakistani engineering students. In *2021 international conference on innovative computing (icic)* (p. 1-10). doi: 10.1109/ICIC53490.2021.9715788

[41] Sharma, M., Joshi, S., Sharma, S., Singh, A., & Gupta, R. (2021). Data mining classification techniques to assign individual personality type and predict job profile. In *2021 9th international conference on reliability, infocom technologies and optimization (trends and future directions) (icrito)* (p. 1-5). doi: 10.1109/ICRITO51393.2021.9596511

[42] Sui, F. M., Chang, J. C., Hsiao, H. C., Chen, S. C., & Chen, D. C. (2018). A study regarding the gap between the industry and academia expectations for college student's employability. In *2018 ieee international conference on industrial engineering and engineering management (ieem)* (p. 1573-1577). doi: 10.1109/IEEM.2018.8607269

[43] Uddin, M. F. (2022). A proposed good fit job candidate algorithm (gfc-a) utilizing big data and career data. In *2022 8th international conference on information technology trends (itt)* (p. 172-176). doi: 10.1109/ITT56123.2022.9863941

[44] Van Huynh, T., Van Nguyen, K., Nguyen, N. L.-T., & Nguyen, A. G.-T. (2020). Job prediction: From deep neural network models to applications. In *2020 rivf international conference on computing and communication technologies (rivf)* (p. 1-6). doi: 10.1109/RIVF48685.2020.9140760

[45] Vassef, S., Toosi, R., & Akhaee, M. A. (2022). Job title prediction from tweets using word embedding and deep neural networks. In *2022 30th international conference on electrical engineering (icee)* (p. 577-581). doi: 10.1109/ICEE55646.2022.9827265

[46] Vinutha, K., & Yogisha, H. K. (2021). Prediction of employability of engineering graduates using

machine learning techniques. In *2021 8th international conference on computing for sustainable global development (indiacom)* (p. 742-745).

[47] Yang, Z., & Cao, S. (2019). Job information crawling, visualization and clustering of job search websites. In *2019 ieee 4th advanced information technology, electronic and automation control conference (iaeac)* (Vol. 1, p. 637-641). doi: 10.1109/IAEAC47372.2019.8997713

[48] Yu, H., & Zhang, Z.-q. (2018). The application of data mining technology in employment analysis of university graduates. In *2018 ieee/acis 17th international conference on computer and information science (icis)* (p. 846-849). doi: 10.1109/ICIS.2018.8466511

[49] Zhou, Q., Liao, F., Ge, L., & Sun, J. (2019). Personalized preference collaborative filtering: Job recommendation for graduates. In *2019 ieee smartworld, ubiquitous intelligence computing, advanced trusted computing, scalable computing communications, cloud big data computing, internet of people and smart city innovation (smartworld/scalcom/uic/atc/cbdcom/iop/sci)* (p. 1055-1062). doi: 10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00203

[50] Zhu, J., Viaud, G., & Hudelot, C. (2021). Improving next-application prediction with deep personalized-attention neural network. In *2021 20th ieee international conference on machine learning and applications (icmla)* (p. 1615-1622). doi: 10.1109/ICMLA52953.2021.00258

# APPENDIX A

# FYP 1 MEETING LOGS

The following pages include the 6 meeting logs for FYP 1.