Development of a Graduate Employability Dashboard for Community Colleges

Muhammad Farhan Bin Hj Azmir, Muhammad Abdul Adib Bin Abd Aziz, Mohd Ruzaimi Bin Mohd Ariffin

Faculty of Industrial Management, Universiti Malaysia Pahang Al-Sultan Abdullah

DOI: https://doi.org/10.47772/IJRISS.2026.1026EDU0361

Received: 20 May 2026; Accepted: 25 May 2026; Published: 19 June 2026

ABSTRACT

Sistem Kajian Pengesanan Graduan (SKPG), also known as the Graduate Tracer Study (GTS), is a structured survey conducted by Malaysia's Ministry of Higher Education (MOHE) to monitor and evaluate the employment outcomes of graduates from higher education institutions. The study plays an important role in examining the alignment between higher education programmes and labour market requirements. It provides valuable information on graduates' employment status, job relevance, career development, employability skills, and satisfaction with the programmes and institutions they attended. The implementation of SKPG involves several important stages, including research design, identification of data sources, selection of the target graduate population, determination of an appropriate survey period, and development of comprehensive survey questionnaires. Data are collected directly from graduates and analysed to support institutional planning, curriculum enhancement, policy formulation, and graduate employability initiatives. This project aims to share international expertise and best practices in implementing and analysing Graduate Tracer Studies with the Jabatan Pendidikan Politeknik dan Kolej Komuniti (JPPKK), Malaysia. In addition, the project proposes the integration of Machine Learning (ML) techniques into SKPG data analysis to improve the prediction of graduate employability outcomes. Through ML-based analysis, educational institutions can identify key employability factors, detect trends in graduate outcomes, and make more informed decisions to align academic programmes with student needs and evolving labour market demands.

Keywords: Graduate Employability; Graduate Tracer Study; Machine Learning; Predictive Analytics; Employability Dashboard; Community College; TVET

INTRODUCTION

Graduate employability has become an increasingly important agenda for higher education institutions, policymakers, industry stakeholders, and workforce development agencies, particularly within a labour market that is becoming more competitive, technology-driven, and rapidly changing. Employers are no longer only looking for academic qualifications; they also expect graduates to demonstrate relevant technical knowledge, practical skills, adaptability, communication abilities, problem-solving capabilities, and readiness to contribute effectively in the workplace. In this context, higher education institutions are expected to ensure that their programmes remain relevant to current industry requirements and future workforce needs. Community colleges play a significant role in supporting national workforce development by providing accessible, practical, and skills-oriented education to diverse groups of learners, including school leavers, working adults, and individuals seeking to improve their technical and vocational competencies. However, ensuring that graduates possess the competencies required to secure relevant and sustainable employment remains a continuing challenge. Graduate employment outcomes are influenced by multiple factors, including field of study, regional labour demand, economic conditions, industrial development, institutional support, work experience, graduate motivation, and individual characteristics such as skills, qualifications, and career readiness.

In Malaysia, the Graduate Tracer Study, locally known as Kajian Pengesanan Graduan (KPG) or Sistem Kajian Pengesanan Graduan (SKPG), serves as an important mechanism for monitoring the employment outcomes of graduates from higher education institutions. The study is designed to collect information directly from graduates regarding their employment status, type of employment, salary level, job relevance, further study involvement, career progression, and satisfaction with their academic programmes and institutions. The data generated from this study can provide meaningful insights into the effectiveness of educational programmes, the relevance of training provided, and the level of alignment between higher education outcomes and labour market needs. Such information is important for institutions to review curriculum content, strengthen industry engagement, improve career services, and develop more targeted employability initiatives. Nevertheless, the effective utilisation of tracer study data requires more than simple reporting. It requires systematic data analysis,

predictive modelling, and user-friendly reporting mechanisms that can translate large volumes of graduate information into meaningful findings that support evidence-based decision-making.

Despite the availability of Graduate Tracer Study data, many institutions continue to face limitations in analysing and interpreting graduate employability patterns in a comprehensive, accurate, and timely manner. Existing approaches to tracking graduate outcomes are often fragmented, manually managed, and largely descriptive in nature. In many cases, the available reports focus mainly on basic statistics, such as the number or percentage of employed graduates, unemployed graduates, and graduates pursuing further studies. Although these descriptive findings are useful, they may not be sufficient to explain why certain graduates experience better employment outcomes than others, which factors contribute most significantly to employability, or which programmes may require additional intervention. This limitation may reduce the ability of institutions to identify emerging employability trends, compare performance across programmes or locations, predict future graduate outcomes, and design strategic interventions based on reliable evidence. This gap highlights the need for a data-driven decision-support tool that can transform raw graduate tracer data into actionable insights, predictive information, and practical recommendations for institutional planning.

Therefore, this project focuses on the development of a Graduate Employability Dashboard for community colleges in Malaysia. The proposed dashboard is designed as an integrated decision-support platform that combines data preprocessing, exploratory data analysis, predictive modelling, and data visualisation to support the analysis and forecasting of graduate employability outcomes. Data preprocessing is necessary to ensure that the tracer study data are complete, consistent, accurate, and suitable for further analysis. Exploratory data analysis will help identify important patterns, relationships, trends, and possible employability gaps among graduates. Predictive modelling will be used to estimate graduate employability outcomes based on selected variables and characteristics within the dataset. Meanwhile, data visualisation will enable users to view and interpret complex findings more easily through charts, indicators, filters, and interactive dashboard elements. By incorporating machine learning techniques and Power BI visualisation, the dashboard is expected to assist administrators, policymakers, career counsellors, programme coordinators, and other relevant stakeholders in making more informed decisions related to curriculum improvement, student support services, resource allocation, industry collaboration, and employability enhancement strategies.

The study uses Graduate Tracer Study data from 88 community colleges under the Jabatan Pendidikan Politeknik dan Kolej Komuniti (JPPKK), with the 2014 graduate cohort selected as the sample dataset. The use of data from multiple community colleges allows the project to examine graduate employability patterns across different institutions, programmes, and geographical areas. The scope of the project includes data cleaning and preprocessing, descriptive and inferential analysis, development and validation of predictive models, and the design of an interactive dashboard. Data cleaning and preprocessing will involve identifying missing values, removing duplicate records, standardising data formats, and preparing relevant variables for analysis. Descriptive analysis will provide an overview of graduate characteristics and employment outcomes, while inferential analysis will examine relationships between selected factors and employability results. The predictive model development process will focus on identifying patterns that may help forecast the likelihood of graduate employment outcomes. The final dashboard will present the results in a clear and accessible format to improve the practical value of graduate employability data.

The significance of this project lies in its contribution to data-driven educational management, institutional planning, and workforce development. By converting raw Graduate Tracer Study data into predictive insights and interactive visualisations, the dashboard can support community colleges in identifying employability gaps, monitoring graduate outcomes, evaluating programme performance, and strengthening institutional strategies. The dashboard may also help institutions identify groups of graduates who may require additional career guidance, training opportunities, industry exposure, or employment support. In addition, the findings may assist JPPKK and other policymakers in formulating more targeted policies and interventions to enhance the employability of community college graduates. Overall, this project demonstrates how data analytics, machine learning, and business intelligence visualisation can be applied to strengthen graduate outcome monitoring, improve institutional responsiveness, and support better alignment between higher education programmes and labour market demands.

LITERATURE REVIEW

Introduction

Graduate employability has been widely recognised as a multidimensional issue shaped by academic preparation, labour market conditions, socio-economic background, institutional support, and individual graduate attributes. In the context of community colleges, graduate employability is particularly important because these institutions play a strategic role in

expanding access to skills-based education and supporting workforce development. Community colleges are expected not only to provide technical and vocational knowledge but also to produce graduates who are capable of meeting the changing demands of industry and the labour market.

This chapter reviews the relevant literature related to graduate employability, with specific attention to Technical and Vocational Education and Training (TVET), graduate tracer studies, national employability strategies, predictive analytics, and dashboard-based decision-support systems. The review also considers the role of the Pelan Strategik Kebolehpasaran Graduan KPT 2021–2025 in strengthening graduate employability in Malaysia. In addition, previous studies on predictive modelling and data visualisation are examined to provide a theoretical and practical foundation for the development of a graduate employability dashboard.

Sistem Kajian Pengesanan Graduan (SKPG)

Sistem Kajian Pengesanan Graduan (SKPG), also known as the Graduate Tracer Study, is a structured survey system implemented by Malaysia's Ministry of Higher Education (MOHE) to monitor the employment outcomes of graduates from higher education institutions. The system functions as an important mechanism for collecting graduate outcome data and assessing the extent to which higher education programmes are aligned with labour market needs.

The implementation of SKPG involves several modes of data collection, including online questionnaires, surveys conducted during graduation ceremonies, and follow-up initiatives managed by institutional alumni offices. Graduates are required to provide information related to their employment or further education status, job search process, and perceptions of how well their academic programmes prepared them for the labour market. This allows institutions to evaluate not only graduate employment outcomes but also the perceived effectiveness of their educational provision.

The data obtained through SKPG are subsequently processed and analysed to identify trends, patterns, and relationships related to graduate employability. Descriptive analysis is commonly used to present the overall employment profile of graduates, while inferential analysis can be applied to examine relationships between selected variables. In more advanced applications, predictive analytics and machine learning techniques may be used to forecast employability outcomes based on academic, demographic, institutional, and labour market-related factors.

Findings from SKPG are usually compiled into reports that provide insights into graduate employability performance, programme relevance, and areas requiring institutional improvement. These reports are valuable for higher education institutions, policymakers, and other stakeholders because they support evidence-based decision-making. For institutions, SKPG findings can guide curriculum review, student support services, career development initiatives, and industry collaboration. For policymakers, the data can inform strategies related to higher education planning, workforce development, and graduate employability enhancement. As an annual and continuous monitoring mechanism, SKPG allows graduate employability trends to be tracked over time. This enables institutions and policymakers to evaluate the long-term impact of educational initiatives and identify emerging challenges in the graduate labour market. Therefore, SKPG serves not only as a reporting tool but also as a strategic instrument for improving the quality, relevance, and responsiveness of higher education in Malaysia.

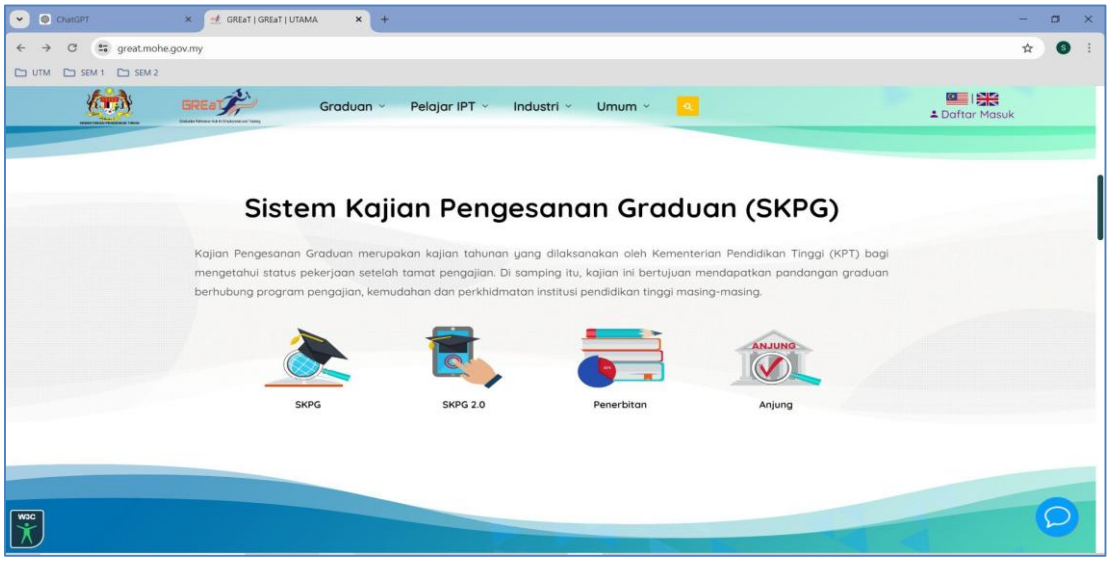


Figure 2.1 Interface of Sistem Kajian Pengesanan Graduan (SKPG). Source: MOHE.

Graduate Employability and Technical and Vocational Education and Training

Technical and Vocational Education and Training (TVET) plays a significant role in enhancing graduate employability by equipping learners with practical, technical, and industry-relevant competencies. Unlike conventional academic pathways, TVET emphasises hands-on learning and occupation-specific skills that are directly applicable to workplace requirements. This makes TVET particularly important in addressing the demands of a dynamic labour market, where practical capabilities, adaptability, and job readiness are increasingly valued by employers.

In Malaysia, TVET has been positioned as a key driver of human capital development and national economic growth. It supports the development of a skilled workforce capable of contributing to industrialisation, technological advancement, and workforce transformation. Through its emphasis on technical skills and practical training, TVET can reduce the mismatch between graduate competencies and employer expectations, thereby improving graduates' transition into employment.

Previous studies and policy discussions have indicated that TVET graduates often demonstrate strong employability outcomes due to their exposure to practical training, industry-based learning, and specialised skill development. Such exposure enables graduates to develop not only technical competencies but also workplace readiness, problem-solving abilities, and familiarity with industry practices. These attributes are essential in strengthening graduates' competitiveness in the labour market.

Despite its potential, TVET continues to face several challenges that may affect its contribution to graduate employability. One of the main challenges is the societal perception that TVET is less prestigious than academic education. This perception may influence students' willingness to pursue TVET pathways and may also affect broader public confidence in vocational education. In addition, issues such as limited funding, uneven quality of facilities, and insufficient industry collaboration can affect the effectiveness of TVET programmes.

Therefore, strengthening TVET requires a comprehensive approach that includes improving public perception, increasing investment in training infrastructure, and enhancing collaboration between educational institutions and industry players. By addressing these issues, TVET can be further established as a credible and valuable educational pathway that supports graduate employability and national workforce development.

Articles and News on Graduate Employability

Graduate employability has become an increasingly important concern, particularly in the post-pandemic labour market, where graduates are expected to demonstrate adaptability, digital competence, industry awareness, and lifelong learning capabilities. Recent discussions in academic literature, policy documents, and media reports suggest that graduate employability is influenced by several key factors, including curriculum relevance, practical training, professional certification, industry engagement, and institutional support mechanisms.

Curriculum design is one of the most important factors in improving graduate employability. Programmes that are continuously reviewed and aligned with industry needs are more likely to produce graduates with relevant knowledge and skills. The integration of industry-based projects, work-based learning, and practical assignments can help students apply theoretical knowledge to real-world contexts. This approach strengthens graduates' ability to respond to workplace challenges and enhances their employability prospects.

In addition to curriculum relevance, employability instruments such as internships, professional certifications, and career development programmes play an important role in preparing graduates for employment. Internships provide students with direct exposure to workplace environments, while professional certifications serve as evidence of specific competencies that are recognised by industry. These elements help bridge the gap between academic preparation and employment requirements.

Furthermore, digital transformation has introduced new employability requirements. The growing demand for digital skills has encouraged initiatives related to digital training, certification, and digital internships. In Malaysia, agencies such as the Malaysia Digital Economy Corporation have contributed to strengthening digital talent development by offering programmes that prepare graduates for opportunities in the digital economy. These initiatives are important because employability is no longer determined solely by academic qualifications, but also by graduates' ability to adapt to technological change.

Overall, the literature and current policy discourse indicate that graduate employability requires a coordinated effort involving curriculum reform, skills development, industry engagement, digital readiness, and evidence-based policymaking. These elements provide an important foundation for the development of data-driven tools, such as employability dashboards, that can support institutions and policymakers in monitoring and improving graduate outcomes.

Predictive Models for Employability Rates

Importance of Predictive Models

Predictive modelling has become an important analytical approach in graduate employability research because it enables institutions to forecast employment outcomes based on historical and institutional data. By identifying the factors associated with graduate employment, predictive models can support educational institutions in designing targeted interventions, improving programme quality, and enhancing student support services.

In the context of graduate tracer studies, predictive modelling allows institutions to move beyond descriptive reporting. While descriptive analysis provides information about employment rates and graduate profiles, predictive analytics can estimate the likelihood of employment based on selected variables such as academic performance, field of study, demographic background, skills, internship experience, and institutional factors. This enables institutions to identify at-risk groups and develop early intervention strategies.

Machine learning has further expanded the potential of predictive modelling in employability studies. Through machine learning techniques, large and complex datasets can be analysed to detect hidden patterns and relationships that may not be easily identified using traditional statistical methods. These models can generate predictive insights that are useful for administrators, policymakers, career counsellors, and programme coordinators.

Therefore, the application of predictive models in graduate employability research is significant because it supports evidence-based decision-making. It enables institutions to better understand employability determinants, forecast graduate outcomes, and align educational strategies with labour market expectations.

Machine Learning Techniques for Graduate Employability Prediction

Machine learning has been increasingly applied in graduate employability studies due to its ability to analyse complex data and generate predictive insights. Employability prediction generally involves the use of algorithms to estimate the likelihood of graduates securing employment based on multiple attributes, including academic achievement, technical skills, soft skills, internship experience, extracurricular involvement, and programme background.

Several machine learning algorithms have been used in previous studies to predict graduate employability. Decision tree models are commonly applied because they provide interpretable outputs and allow stakeholders to understand the relationship between predictor variables and employability outcomes. Support Vector Machines have also been used due to their ability to classify data effectively, particularly when dealing with complex patterns. Neural networks, on the other hand, are useful for modelling non-linear relationships, although they may require larger datasets and more careful interpretation.

Other commonly used algorithms include Random Forest, K-Nearest Neighbours, and automated machine learning approaches. Random Forest is useful because it combines multiple decision trees to improve prediction accuracy and reduce overfitting. K-Nearest Neighbours classifies cases based on similarity patterns, while automated machine learning supports model selection and hyperparameter optimisation. These techniques can improve the efficiency and accuracy of employability prediction models.

Previous studies have demonstrated the relevance of machine learning in predicting employability among graduates from different fields, including information technology, management, and engineering. These studies highlight that employability is influenced by a combination of academic, technical, experiential, and personal factors. For example, internship experience, project involvement, technical competence, and continuous learning have been identified as important contributors to employment outcomes.

In the Malaysian context, the use of data mining and machine learning in graduate employability research provides an opportunity to strengthen the analysis of Graduate Tracer Study data. By applying predictive models to tracer study datasets, institutions can identify key employability indicators and develop more targeted strategies to support graduates. Such

models can also be integrated into dashboard systems to present predictive findings in a more accessible and actionable format.

Therefore, machine learning offers strong potential to enhance graduate employability analysis by improving prediction accuracy, supporting institutional decision-making, and enabling more proactive employability interventions. This provides a relevant foundation for the development of a graduate employability dashboard that integrates data visualisation and predictive analytics for community colleges in Malaysia.

Machine Learning Algorithm	Zheng (2023)	ElSharkawy et al. (2022)	Shahriyar et al. (2022)	Vinutha and Yogisha (2021)	Raman and Pramod (2021)	Sapaat et al. (2011)
Decision Tree	/	/	/		/	
Gaussian Naïve Bayes		/	/		/	
Logistic Regression		/		/	/	
Random Forest		/		/	/	
Support Vector Machine		/		/		
J48 Decision Tree						/
Naïve Bayes				/		
Bayesian Classifier				/		/
Artificial Neural Network				/		
Gradient Boosting				/	/	
Xgboost				/		
Bagging Classifier					/	
Extra Trees Classifier					/	
KNN			/		/	
RIdgeClassifierSVC					/	
AutoML			/			
Multilayer Perceptron	highest accuracy when compared to other algorithms		/			
BernoulliNB					/	
Ada Boost					/	

Figure 2.3 Comparison machine learning technique from various literature

Performance Evaluation

With the increase in the adoption rate of machine learning algorithms in multiple sectors, the need for accurate measurement and assessment is imperative, especially when classifiers are applied to real world applications. Determining which are the most appropriate evaluation metrics to effectively assess and evaluate the performance of a binary, multi-class and multi-labelled classifier needs to be further understood. Another significant challenge impacting research is that results from models that are similar in nature cannot be adequately compared if the criteria for the measurement and evaluation of these models are not standardized. This review paper aims at highlighting the various evaluation metrics being applied in research and the non-standardization of evaluation metrics to measure the classification results of the model. Although Accuracy, Precision, Recall and F1-Score are the most applied evaluation metrics, there are certain limitations when considering these metrics in isolation. Other metrics such as ROC\AUC and Kappa statistics have proven to provide additional insightful into the effectiveness of an algorithms adequacy and should also be considered when evaluating the effectiveness of binary, multi-class and multi-labelled classifiers. The adoption of a standardized and consistent evaluation methodology should be explored as an area of future work.

Performance evaluation is a crucial part of the machine learning pipeline. It helps determine how well a model is performing and whether it's making progress. Below are some common performance metrics used in machine learning:

Applications and Outcomes

These ML techniques have been applied across various datasets, including academic records, demographic information, and external labor market data. The outcomes of these studies highlight the potential of ML models to provide actionable insights for educational institutions and policymakers. For instance, integrating ML predictions into academic advising systems can help tailor educational programs to better meet the needs of students and the labor market.

Dashboards for Graduate Employability

Dashboards are interactive tools that visualize data, making it easier for users to understand complex information. For

graduate employability, dashboards can display metrics such as employment rates, average salaries, and skill gaps, providing valuable insights for administrators, career counselors, and students. Effective employability dashboards typically include:

(a) Use of charts, graphs, and heatmaps to illustrate trends and patterns in graduate employability.

(b) Options to filter data by various criteria such as program, graduation year, and demographic factors, allowing users to customize their analysis.

(c) Integration of predictive models to forecast future employability outcomes trends.

(d) An intuitive design that allows users to easily navigate and interpret the data, making the dashboard accessible to a wide range of users.

Dashboard Development

Dashboard development is the process of creating a visual display of data that helps users monitor, analyse, and explore information. Dashboards can be used for various purposes, such as business intelligence, project management, performance tracking, and more.

There are different steps and tools involved in dashboard development, depending on the data source, the audience, the design, and the functionality of the dashboard. some general steps to follow for dashboard development are as Table 1.

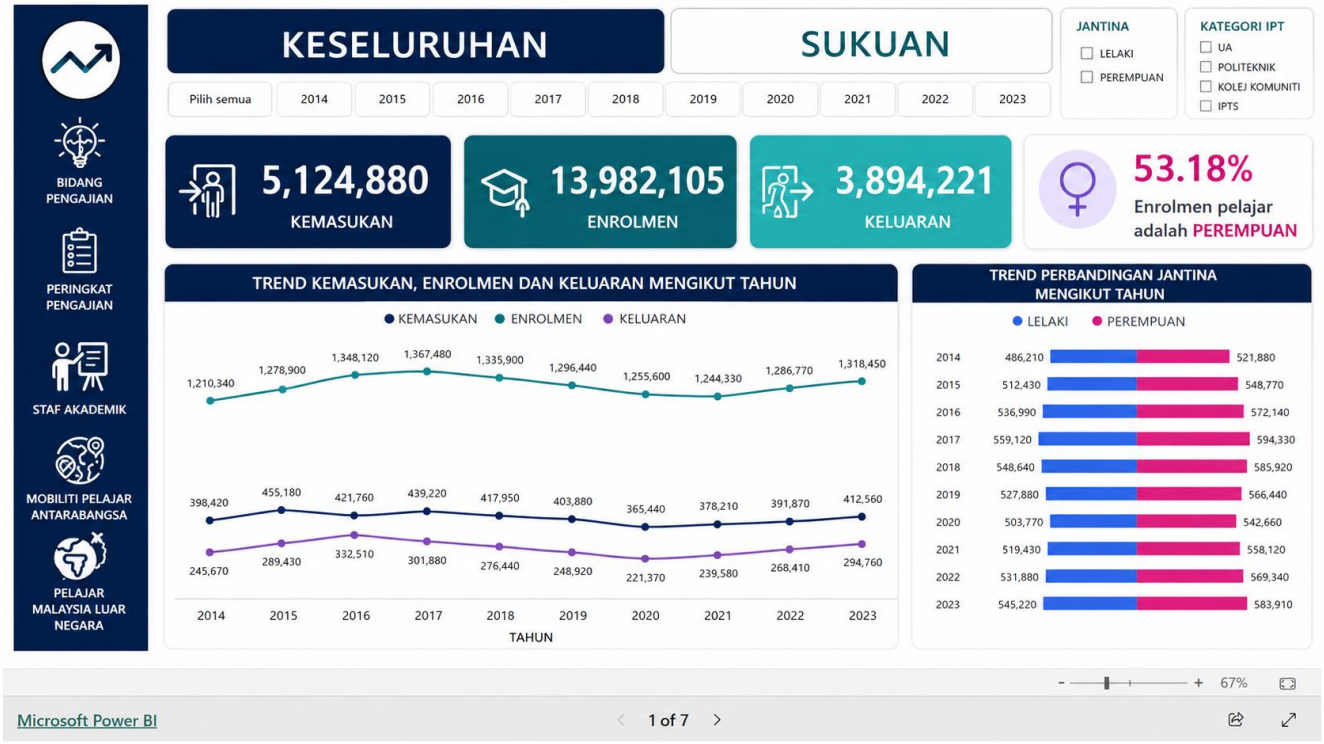
Table 2.1General steps for dashboard development

No.	Steps	Questions
1	Define the goal and scope of the dashboard:	- What is the purpose of the dashboard? - Who will use it? - What questions will it answer? - What data will it show?
2.	Collect and prepare the data:	- Where will the data come from? - How will it be cleaned, organized, and transformed? - What metrics and dimensions will be used?
3.	Choose a dashboard tool:	- What software or platform will be used to create the dashboard? - What features and capabilities does it offer? - How easy is it to use and maintain?
4.	Design the dashboard layout: How will the dashboard look like?	- What visual elements will be used, such as charts, tables, maps, etc.? - How will they be arranged and formatted? - How will the dashboard support interactivity and navigation?

Case Studies

The Ministry of Higher Education Malaysia's Dashboard Statistic

The Ministry of Higher Education Malaysia's Dashboard Statistic provides comprehensive and interactive data on various aspects of higher education. This includes statistics on student enrolment, graduate employability, academic programs, and institutional performance. The dashboard aims to offer stakeholders, including policymakers, educators, and the public, a clear and detailed overview of the higher education landscape in Malaysia. It is designed to support data-driven decision-making and improve the transparency and effectiveness of educational initiatives.



Several improvements could be considered like improving data visualization with dynamic charts and graphs would make trends and comparisons clearer. Implementing real-time updates would keep the data current. Offering user customization options for personalized views and reports could increase engagement. Providing a more detailed analysis of factors influencing employability and sector-specific trends would allow user-driven experience.

Figure 2.4 The Ministry of Higher Education Malaysia's Dashboard Statistic (2024). Source: MOHE. Retrieved from https://www.mohe.gov.my/en/broadcast/dashboard-statistic

Graduate Employability Analysis UMPSA

Figure 2.5 shows a dashboard for Graduate Employability Analysis UMPSA provides a comprehensive overview of factor employment metrics among recent graduates. It includes descriptive statistics on gender distribution, modes of study, racial demographics, geographical locations of employment, distribution across various fields, age by program, and levels of certification achieved. This data visualization tool enables stakeholders to identify trends, disparities, and opportunities within the graduate job market.

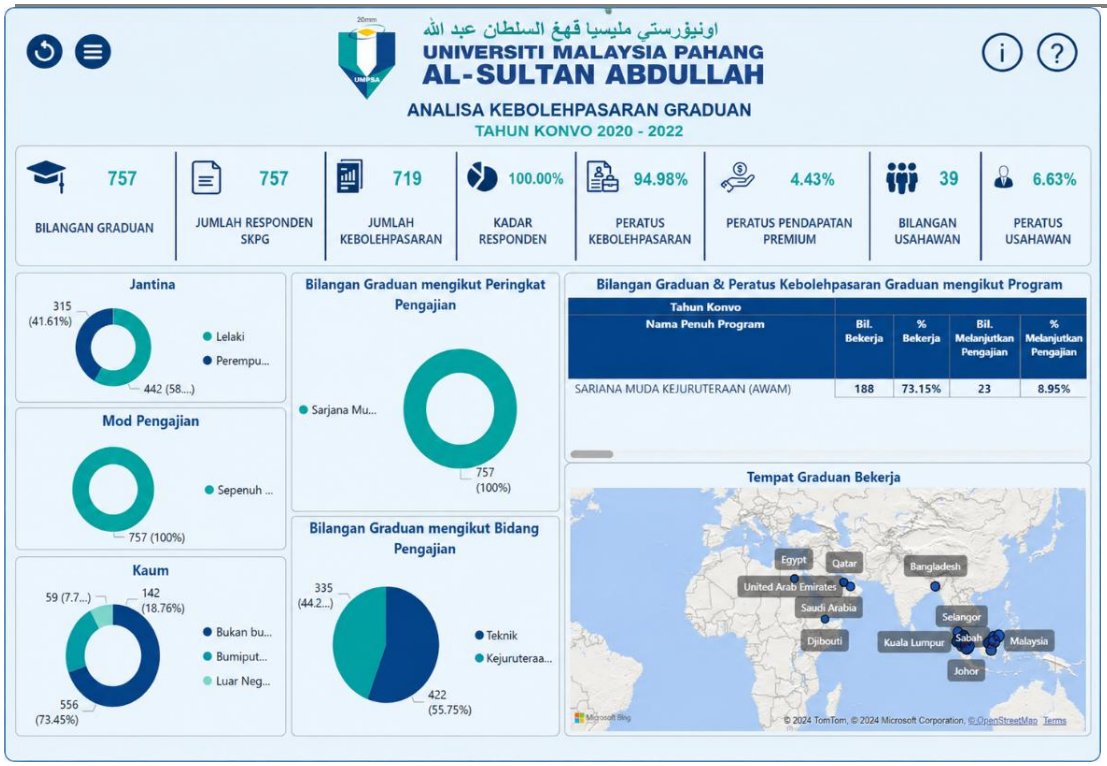


Figure 2.5 UMPSA Analisa Kebolehpasaran Graduan Dashboard (2023)

The graduate employability dashboard is designed as an intuitive tool that offers interactive filters for users to examine specific data subsets, such as gender, race, and location. It incorporates trend analysis through trend lines and bar graphs to visualize changes in employability metrics over time. Comparative views are available to assess differences in employability outcomes by program or certificate level. Geographical mapping is utilized to depict the distribution of graduates' workplaces, while field distribution is represented with pie charts or heat maps for a clear visual breakdown. Additionally, age analysis segments data by program to reveal which age groups fare better in certain fields. Most importantly, the dashboard provides actionable insights that can guide educational institutions, policymakers, and graduates in making informed decisions to improve employability outcomes.

The Carnegie Mellon University Post-Graduation Outcomes Dashboard

The Carnegie Mellon University Post-Graduation Outcomes Dashboard provides detailed insights into the career paths of graduates. It includes data on hiring companies, graduate schools, starting salaries, and geographic locations of recent alumni. Users can filter information by academic department and major, and access interactive elements such as charts and maps for a comprehensive view. This tool supports data-driven career planning and institutional analysis.

To improve Carnegie Mellon University's Post-Graduation Outcomes dashboard, several improvements could be made. More interactive and visually appealing charts and graphs would help users better understand trends and data. Integrating real-time data updates would ensure the information remains current. Extend to mobile compatibility would make the dashboard more accessible.

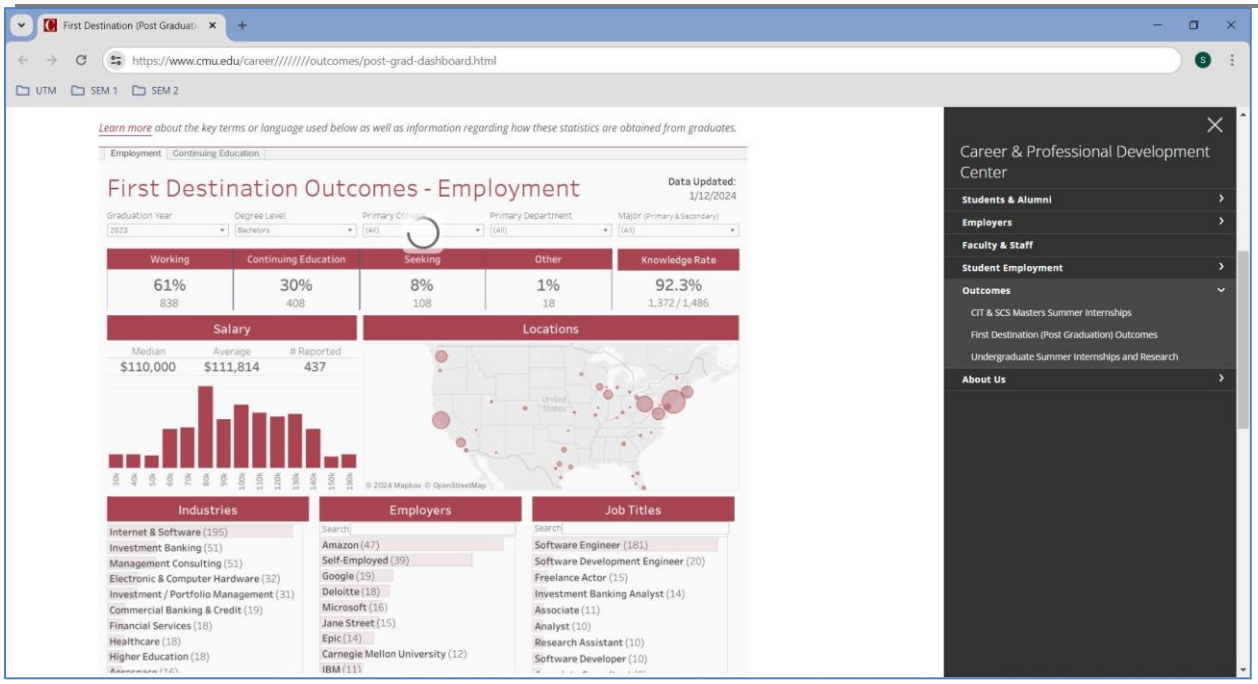


Figure 2.6 The Carnegie Mellon University Post-Graduation Outcomes Dashboard 2024. Source: National Employment Database. Retrieved from https://www.cmu.edu/career/////////outcomes/post-grad-dashboard.html

The Ministry of Higher Education Malaysia's Graduate Tracer Study Report 2022

Figure 2.8, 2.9 and 2.10 is a sample of static report from The Ministry of Higher Education Malaysia's Graduate Tracer Study Report 2022. Static reports present data in a fixed format, often as tables or charts within a report. While they allow clear comparisons over time, they lack interactivity. In contrast, machine learning dashboards offer dynamic data visualization, allowing users to interact with real-time charts and graphs. These dashboards integrate advanced analytics, such as predictive insights and anomaly detection. Users can customize views, track data in real-time, and make informed decisions. By leveraging machine learning, educational institutions enhance reporting and proactive management of graduate employability.

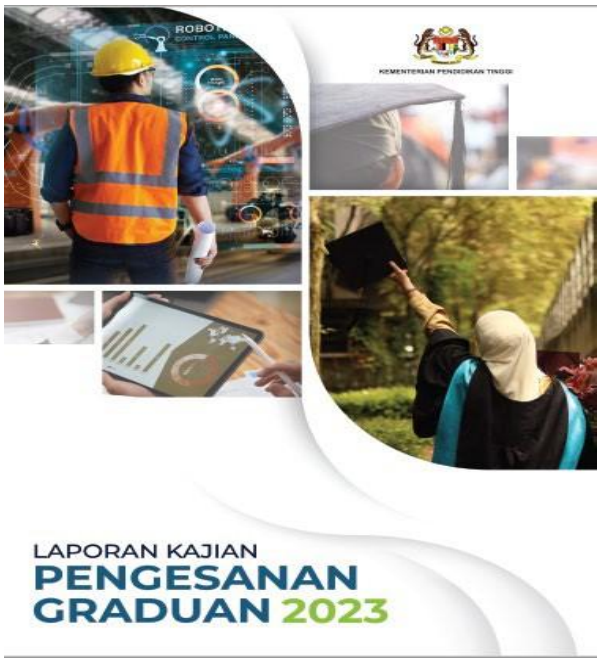


Figure 2.7 Cover The Ministry of Higher Education Malaysia's Graduate Tracer Study Report 2022 by KPT, Laporan Kajian Pengesanan Graduan 2023

Using dashboard machine learning by integrating several advanced techniques can make information more dynamic. Interactive visualizations allow users to explore data by filtering and drilling down, with tools like Tableau or Power BI enhancing these interactions. Predictive analytics can be incorporated to provide forecasts and insights based on historical data, while anomaly detection algorithms can automatically highlight unusual patterns. Clustering and segmentation techniques help visualize different data groups more clearly. Dynamic filtering based on machine learning insights can refine data views, and natural language processing (NLP) can generate automated summaries or explanations alongside charts and tables. Real-time data updates ensure that dashboards reflect the most current information, and user personalization features tailor the experience to individual preferences. These strategies collectively enhance the interactivity, insightfulness, and overall effectiveness of your dashboard.

Tools and Platforms

Microsoft Excel

Excel is a versatile tool for data preprocessing due to its ability to handle various tasks such as cleaning, transforming, and organizing data. Excel provides a comprehensive set of tools and functions that can streamline the data preprocessing workflow, making it suitable for various data cleaning and transformation tasks before further analysis or modelling in other tools like Python.

Python

Python is a versatile, high-level programming language created by Guido van Rossum and first released in 1991. Known for its simplicity and readability, Python is beginner-friendly yet powerful enough for complex applications. It is an interpreted language, meaning code is executed line by line, which facilitates debugging and development. Python is cross-platform, running seamlessly on Windows, macOS, and Linux, and boasts an extensive ecosystem of libraries for diverse fields like data science, machine learning, web development, automation, and scientific computing. Popular libraries include NumPy and pandas for data analysis, TensorFlow and PyTorch for machine learning, Django and Flask for web development, and Matplotlib and Seaborn for data visualization. With its massive community support, Python offers abundant resources and tutorials, making it accessible to developers of all levels. Its combination of simplicity, versatility, and a vast array of tools has made Python one of the most popular programming languages worldwide.

Microsoft Power BI

Microsoft Power BI is a powerful business intelligence tool used to develop interactive dashboards and perform data analytics. Here's a description of how Power BI is used in developing dashboards for data analytics: Microsoft Power BI is a suite of business analytics tools that enables organizations to visualize and share insights from their data. It connects to a wide range of data sources, transforms raw data into meaningful information, and offers rich interactive visualizations and business intelligence capabilities. Microsoft Power BI empowers organizations to leverage their data for informed decision-making by creating visually compelling and interactive dashboards that deliver actionable insights across the enterprise.

No.	Steps	Questions
5.	Build and test the dashboard:	- How will the dashboard be implemented and connected to the data source? - How will the dashboard be tested for accuracy, functionality, and usability? - How will the dashboard be deployed and shared with the users?
6.	Evaluate and improve the dashboard:	- How will the dashboard be monitored and maintained? - How will the feedback from the users be collected

		and incorporated? - How will the dashboard be updated and improved over time?

RESEARCH METHODOLOGY

Introduction

This chapter presents the methodological approach adopted in developing the Community College Graduate Employability Dashboard. The study was designed to transform graduate tracer data into meaningful analytical outputs through data preprocessing, exploratory analysis, predictive modelling, and dashboard visualisation. The overall methodology integrates data analytics and machine learning techniques with business intelligence tools to support evidence-based decision-making related to graduate employability.

The development process consisted of three main components: exploratory data analysis, predictive model development, and interactive dashboard design. Data preprocessing and analysis were conducted using Microsoft Excel and Python, while Power BI was used to develop the dashboard interface. These tools were selected because they support data cleaning, statistical exploration, model development, and interactive visualisation, which are essential for producing reliable and accessible graduate employability insights.

This chapter explains each methodological stage in detail, including the research framework, data preparation process, model development procedure, and dashboard development process. The explanation provides a clear structure for understanding how the graduate employability dashboard was designed, developed, and evaluated.

Research Framework

The research framework was developed to guide the systematic development of the Community College Graduate Employability Dashboard. As illustrated in Figure 3.1, the methodology consists of five main phases: problem formulation, data gathering, data preprocessing, model development, and dashboard development. These phases were arranged sequentially to ensure that the project was conducted in a structured and logical manner.

The first phase involved problem formulation, where the research issue, objectives, and expected outcomes were identified. This phase focused on understanding the need for a data-driven tool that can assist community colleges and relevant stakeholders in analysing and predicting graduate employability outcomes.

The second phase was data gathering, which involved obtaining the Graduate Tracer Study dataset related to community college graduates. This dataset provided the foundation for subsequent analysis, as it contained information relevant to employment status, graduate background, programme details, and other employability-related variables.

The third phase involved data preprocessing. In this phase, the dataset was cleaned, organised, and transformed to ensure that it was suitable for further analysis and modelling. This included checking missing values, removing inconsistencies, standardising variables, and preparing the data for exploratory data analysis and machine learning processes.

The fourth phase focused on model development. Predictive models were developed to forecast graduate employability outcomes based on selected variables from the dataset. Machine learning techniques were applied to identify patterns and relationships that could support the prediction of graduate employment status. The models were then evaluated to determine their accuracy and suitability for integration into the dashboard.

The final phase was dashboard development. The results from the data analysis and predictive modelling process were visualised through an interactive Power BI dashboard. The dashboard was designed to present key employability indicators, trends, and model outputs in a user-friendly format. This allows administrators, policymakers, and other stakeholders to interpret graduate employability data more effectively and use the insights for decision-making.

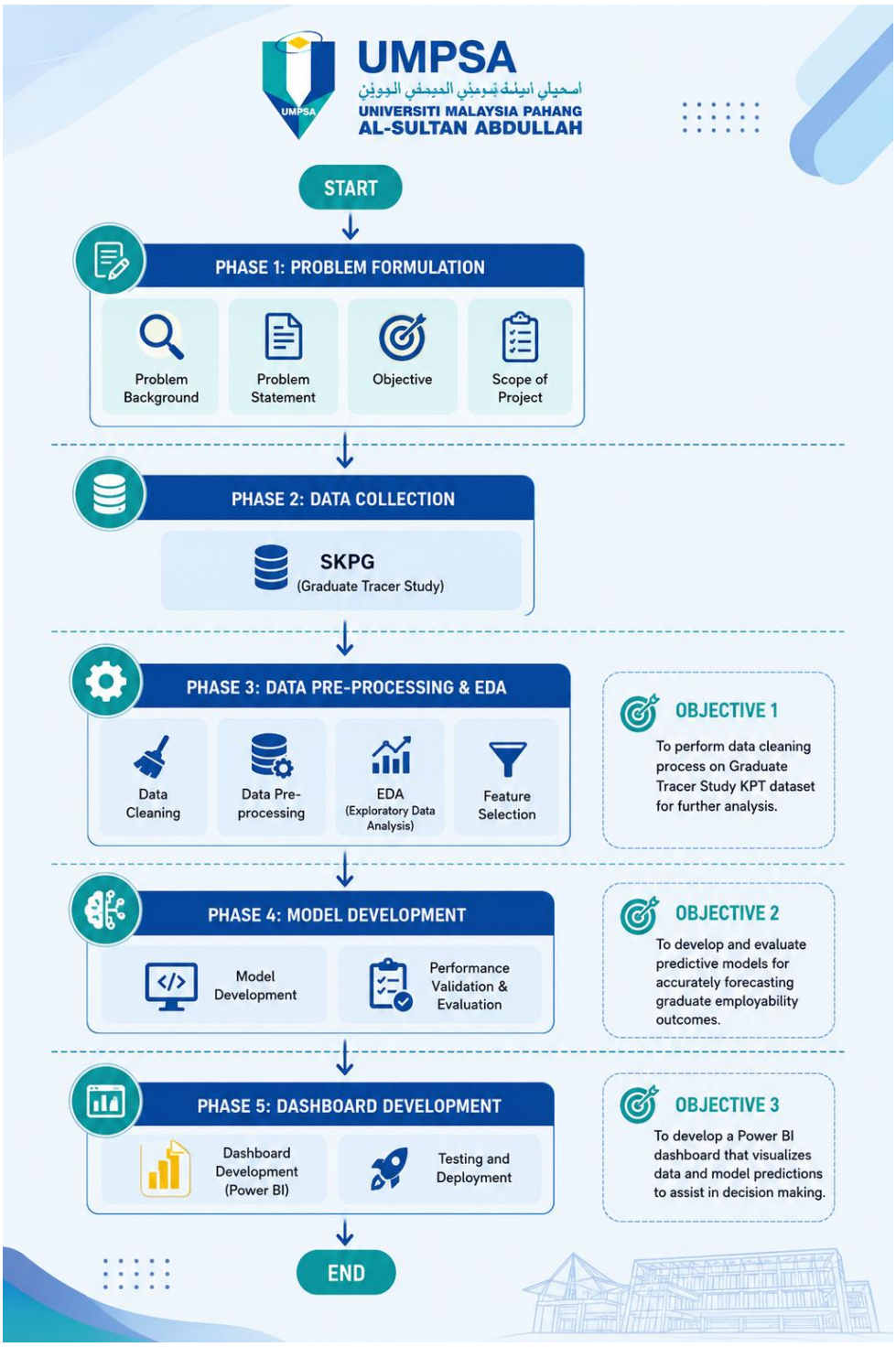


Figure 3.1 Research methodology process flow of College Community Graduate Employability Dashboard Development

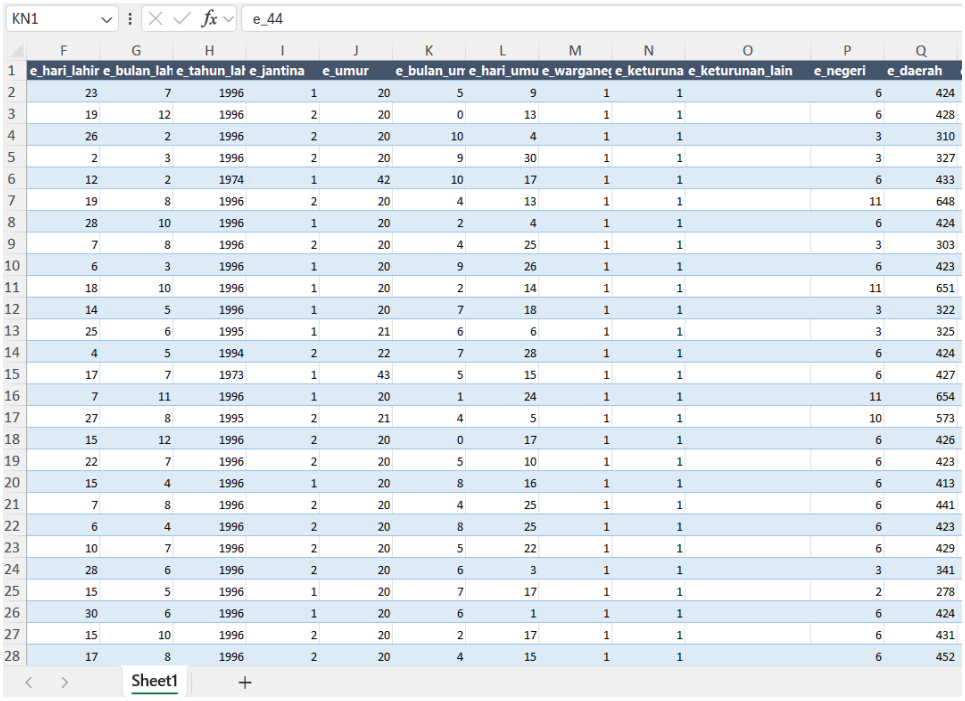
Problem Formulation

The research begins with problem formulation, which involves understanding the background, defining the problem statement, setting objectives, and outlining the scope of the project. This phase includes problem formulation and a literature review, as covered in Chapters 1 and 2. Chapter 1 outlines the problem statement and background, highlighting the importance of data cleaning for the Graduate Tracer Study KPT dataset and the challenges in accurately forecasting graduate employability outcomes. Chapter 2 reviews existing literature on predictive modeling and data visualization, discussing relevant studies, methodologies, techniques, and the importance of visual tools like Power BI for decision-making. Various resources and research on predictive modeling for employability outcomes were examined and used as research benchmarks.

Data Definition and Collection

In this phase, data is collected from the Graduate Tracer Study (SKPG) conducted by JPPKK. The data includes information on graduates' employment status, field of study, and other relevant variables. Data was collected from various sources including community college records, employment databases, and surveys conducted among graduates. The collected data included variables such as graduate demographics, academic performance, job placements, industry sectors, and job roles.

The dataset encompasses several attribute characteristics, including data collected from the years 2014. It contains a total of 3,118 instances and 357 attributes. The attributes include various types such as nominal, scale, and ordinal. It is important to note that the dataset has missing values, which may impact the analysis and require appropriate handling.



KN1 | e_44

	F	G	H	I	J	K	L	M	N	O	P	Q
1	e_hari_lahir	e_bulan_lah	e_tahun_lah	e_jantina	e_umur	e_bulan_um	e_hari_umu	e_warganeg	e_keturuna	e_keturunan_lain	e_negeri	e_daerah
2	23	7	1996	1	20	5	9	1	1		6	424
3	19	12	1996	2	20	0	13	1	1		6	428
4	26	2	1996	2	20	10	4	1	1		3	310
5	2	3	1996	2	20	9	30	1	1		3	327
6	12	2	1974	1	42	10	17	1	1		6	433
7	19	8	1996	2	20	4	13	1	1		11	648
8	28	10	1996	1	20	2	4	1	1		6	424
9	7	8	1996	2	20	4	25	1	1		3	303
10	6	3	1996	1	20	9	26	1	1		6	423
11	18	10	1996	1	20	2	14	1	1		11	651
12	14	5	1996	1	20	7	18	1	1		3	322
13	25	6	1995	1	21	6	6	1	1		3	325
14	4	5	1994	2	22	7	28	1	1		6	424
15	17	7	1973	1	43	5	15	1	1		6	427
16	7	11	1996	1	20	1	24	1	1		11	654
17	27	8	1995	2	21	4	5	1	1		10	573
18	15	12	1996	2	20	0	17	1	1		6	426
19	22	7	1996	2	20	5	10	1	1		6	423
20	15	4	1996	1	20	8	16	1	1		6	413
21	7	8	1996	2	20	4	25	1	1		6	441
22	6	4	1996	2	20	8	25	1	1		6	423
23	10	7	1996	2	20	5	22	1	1		6	429
24	28	6	1996	2	20	6	3	1	1		3	341
25	15	5	1996	1	20	7	17	1	1		2	278
26	30	6	1996	1	20	6	1	1	1		6	424
27	15	10	1996	2	20	2	17	1	1		6	431
28	17	8	1996	2	20	4	15	1	1		6	452

Sheet1

Figure 3.2 Screenshot of an Excerpt from Raw Dataset of College Community Graduate Employability store in excel file

Figure 3.2 presents a screenshot of an excerpt from the raw dataset on College Community Graduate Employability, which is stored in an Excel file. The dataset comprises 357 attributes covering various aspects of graduate employability, including demographic details, academic performance, extracurricular involvement, curriculum relevance, career guidance experiences, skill development, and employment-related factors. To provide an overview of key attributes, Table 3.1 lists a selection of key attributes along with their descriptions.

Table 3.1 Description of Graduate Employability Dataset Attributes

No.	Attributes	Descriptions
1	e_hari_lahir	Tarikh Lahir (Hari)
2	e_bulan_lahir	Tarikh Lahir (Bulan)

3	e_tahun_lahir	Tarikh Lahir (Tahun)
4	e_jantina	Jantina
5	e_umur	Umur
6	e_bulan_umur	Bulan Umur
7	e_hari_umur	Hari Umur
8	e_warganegara	Warganegara
9	e_keturunan	Keturunan
10	e_keturunan_lain	Keturunan Lain
11	e_negeri	Negeri Bermastautin
12	e_daerah	Daerah

Data Preprocessing and EDA

In this phase, the datasets will undergo data cleaning and preprocessing to prepare them for machine learning. Effective preprocessing is crucial for transforming raw data into a clean format, which significantly impacts the quality of the outcomes.

Data cleaning, data pre-processing and EDA

Data preprocessing is essential for preparing raw data for meaningful analysis. This step includes several crucial tasks to ensure the dataset is clean and ready for modeling. First, handling missing values involves identifying and addressing any gaps or incomplete entries, which may be done through imputation or removal of affected records. Next, encoding categorical variables transforms non-numeric data into a numerical format suitable for analysis, using methods such as one-hot or label encoding. Normalizing numerical features is also important, as it standardizes values to a common scale, which is vital for algorithms sensitive to data scale. After preprocessing, Exploratory Data Analysis (EDA) is performed to understand the data's distribution, detect anomalies, and explore relationships between variables. Table 3.2 illustrates a systematic approach to preprocessing the graduate employability dataset, ensuring that it is clean, consistent, and ready for effective analysis and modeling.

Table 3.2: Pre-processing steps, actions, and expected output

Pre-processing Steps	**Actions**	**Expected Output**
Data Cleaning	Remove duplicates	Clean dataset without redundant entries
Data Cleaning	Handle missing values (e.g., impute, remove)	Complete dataset without missing values
Data Cleaning	Correct data entry errors	Accurate and consistent data
Data Transformation	Encode categorical variables (e.g., one-hot encoding)	Numerical representation of categorical variables
Data Transformation	Normalize/Standardize numerical features	Scaled data with mean = 0 and standard deviation = 1
Data Transformation	Discretize continuous variables if needed	Discrete bins for continuous data
Data Reduction	Feature selection (e.g., remove irrelevant features)	Reduced dataset with only relevant features

Data Integration	Combine data from multiple sources	Integrated dataset with all necessary information.
Data Formatting	Ensure consistent data formats (e.g., dates, numbers)	Uniformly formatted dataset
Data Formatting	Rename columns for clarity	Clear and descriptive column names
Data Splitting	Split data into training, validation, and test sets	Separate datasets for model development and evaluation

Feature Selection

After initial cleaning, Information Gain will be employed to select the most relevant features from the dataset. Information Gain measures how much information a feature provides about the target variable, helping to identify features that significantly reduce uncertainty.This process involves calculating Information Gain for each feature and selecting those with the highest values for further analysis. Features with higher Information Gain are more informative and will be retained, while less significant features will be discarded. Table 3.3 explain step in feature selection process.

Table 3.3 Feature selection process

Feature Selection Process	**Process Explanation**
Calculate Entropy for the Entire Dataset	• Determine the initial entropy of the dataset before any feature-based splitting.
Compute Information Gain for Each Feature	• For each feature, calculate the entropy of the subsets created by splitting the dataset based on the feature's values. • Compute the Information Gain by subtracting the weighted average entropy of these subsets from the initial entropy.
Rank Features	• Rank features based on their Information Gain values. Features with higher Information Gain are considered more informative.
Select Top Features	• Choose a subset of top-ranked features with the highest Information Gain for use in building predictive models.

Information Gain is a key concept used in machine learning for feature selection, particularly in the context of decision trees. It measures how much information a feature contributes to reducing uncertainty or entropy regarding the target variable.

Entropy is a measure of uncertainty or randomness in a dataset. It quantifies the impurity or disorder in the data.

Information Gain measures the reduction in entropy (uncertainty) when a dataset is split based on a particular feature. It indicates how much a feature helps in making the data more homogeneous or pure.

A high Information Gain indicates that the feature significantly reduces uncertainty about the target variable, making it a good candidate for inclusion in the model. A low Information Gain suggests that the feature provides little information about the target variable and may be less useful for prediction.

a) Initial Entropy Calculation: Compute the entropy of the target variable (employment status) for the entire dataset.

b) Feature-based Splitting: For "Field of Study", split the dataset into subsets where each subset corresponds to a different

field. Compute the entropy for each subset and the weighted average entropy of these subsets.

c) Information Gain Calculation: Subtract the weighted average entropy from the initial entropy to get the Information Gain for "Field of Study".

d) Repeat for Other Features: Perform the same calculation for features like "GPA".

e) Select Features: If "Field of Study" has a higher Information Gain compared to "GPA", it indicates that "Field of Study" provides more useful information for predicting employability.

Information Gain is used in algorithms to build decision trees. The algorithm chooses features with the highest Information Gain for splitting nodes, aiming to create the most informative and concise tree structure.

Model Development

The next stage involved developing predictive models to forecast employability outcomes based on the available data. Predictive models are developed using Decision Tree and Random Forest algorithms. These models are trained on the pre-processed data to predict graduate employability outcomes. The performance of these models is validated and evaluated using appropriate metrics to ensure their accuracy and reliability.

With the feature extraction process complete, the next phase involves developing predictive models to forecast graduate employability outcomes. This process utilizes machine learning algorithms implemented through the Scikit-Learn library. Initially, standalone classifiers are built, starting with the Decision Tree algorithm. A Decision Tree model is trained on the pre-processed data, where the algorithm learns to make predictions by creating a flowchart-like structure that splits the data based on feature values to best classify the employability outcomes. Following this, a Random Forest classifier is developed. Unlike a single Decision Tree, a Random Forest consists of multiple Decision Trees trained on various subsets of the data. This ensemble method aggregates the predictions from each tree to provide a final classification, enhancing overall accuracy and robustness.

After training the standalone classifiers, the performance of each model is evaluated using a validation set. Key performance metrics, such as accuracy, precision, recall, and F1-score, are used to assess how well the models predict employability outcomes. To further refine the models, hyperparameter tuning is performed. This involves using Grid Search techniques to identify the optimal parameters for each model, such as the depth of the trees in Decision Tree and the number of trees or features in the Random Forest. The best parameters are then applied to retrain the models, which are subsequently tested on a separate test set to validate their performance.

Development of Dashboard

The final stage involved developing dashboard using Power BI to visualize the data and prediction results. This dashboard provides stakeholders with an intuitive interface to explore employability trends and model predictions. The final phase involves developing a dashboard using Power BI. The dashboard integrates data visualization and predictive analytics to provide a comprehensive tool for stakeholders to make informed decisions. It includes visual representations of data trends, model predictions, and other relevant insights.

The dashboard visualization will examine the relationships between graduate employability metrics and various factors, aiming to uncover patterns, correlations, and trends in how these factors influence employability outcomes. Interactive features, including filters, will be integrated to enable users to focus on specific metrics and enhance their exploration of the data. This approach will improve user experience and engagement. The visual elements such as charts, graphs, and tables will be strategically incorporated to effectively communicate insights derived from the employability analysis. Figure 3.2 illustrates the initial dashboards created using Power BI.

Figure 3.3: Sample dashboard for College Community Graduate Employability Dashboard

DISCUSSION

The primary objective of this project is to construct a Graduates Employability Model using a classification task. The dataset used for this study is sourced from the Kajian Pengesanan Graduan KPT (2014), comprising 3,118 responses from graduates of Community College who completed their convocation during the specified period. The dataset includes a wide range of attributes such as demographic information, academic performance, and employability status, which are pivotal for analyzing graduates' employability trends.

Data Exploration

As discussed in chapter 3, the data collection process involved gathering information from the Kajian Pengesanan Graduan KPT database with additional data from the district table, including district_code, district_name, district_name_1, state, latitude, longitude, and the IPT table, which contains IPT_ID and IPT_NAME.

Table 4.1: Information of dataset

Number of Attributes	3,118
Number of Instances	357
Missing Value	Yes
Attribute Characteristics	Nominal, Scale, Ordinal
Data Collected	2014
Programs	24
Community Colleges	68 across Malaysia

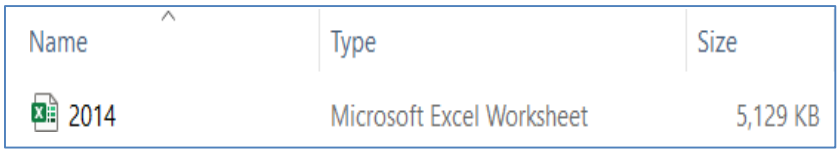


Figure 4.1: Main dataset file that will be used in this project

```
# Load the dataset
# Replace 'file_path' with the actual path to your CSV/Excel file
data = pd.read_excel("2014.xlsx")
print("\nDataset Information:")
data.info()
```

```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3118 entries, 0 to 3117
Columns: 357 entries, tahun_data to e_status
dtypes: float64(158), int64(146), object(53)
memory usage: 8.5+ MB
```

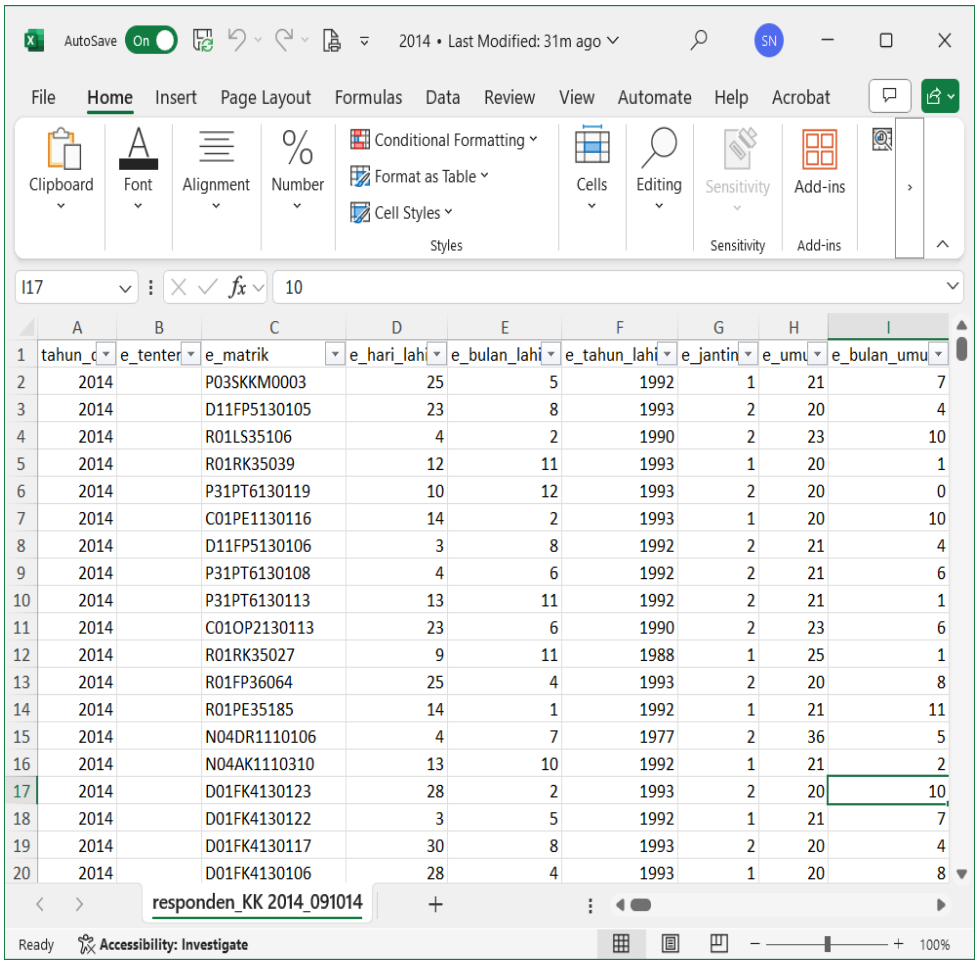
```
print(data.head(5))
```

```
   tahun_data  e_tentera      e_matrik  e_hari_lahir  e_bulan_lahir  \
0        2014        NaN   P03SKKM0003            25              5
1        2014        NaN  D11FP5130105            23              8
2        2014        NaN    R01LS35106             4              2
3        2014        NaN    R01RK35039            12             11
4        2014        NaN  P31PT6130119            10             12

   e_tahun_lahir  e_jantina  e_umur  e_bulan_umur  e_hari_umur  ...  \
0           1992          1      21             7            7  ...
1           1993          2      20             4            9  ...
2           1990          2      23            10           25  ...
3           1993          1      20             1           19  ...
4           1993          2      20             0           22  ...
```

Figure 4.2 Using Python to view dataset

Table 4.2 Number of attribute and instances for each file

File Name	Number of Attribute	Number of Instances
2014	357	3118
IPT	2	47
district	6	804



	A	B	C	D	E	F	G	H	I
1	tahun_c	e_tenter	e_matrik	e_hari_lahi	e_bulan_lahi	e_tahun_lahi	e_jantin	e_umu	e_bulan_umu
2	2014		P03SKKM0003	25	5	1992	1	21	7
3	2014		D11FP5130105	23	8	1993	2	20	4
4	2014		R01LS35106	4	2	1990	2	23	10
5	2014		R01RK35039	12	11	1993	1	20	1
6	2014		P31PT6130119	10	12	1993	2	20	0
7	2014		C01PE1130116	14	2	1993	1	20	10
8	2014		D11FP5130106	3	8	1992	2	21	4
9	2014		P31PT6130108	4	6	1992	2	21	6
10	2014		P31PT6130113	13	11	1992	2	21	1
11	2014		C01OP2130113	23	6	1990	2	23	6
12	2014		R01RK35027	9	11	1988	1	25	1
13	2014		R01FP36064	25	4	1993	2	20	8
14	2014		R01PE35185	14	1	1992	1	21	11
15	2014		N04DR1110106	4	7	1977	2	36	5
16	2014		N04AK1110310	13	10	1992	1	21	2
17	2014		D01FK4130123	28	2	1993	2	20	10
18	2014		D01FK4130122	3	5	1992	1	21	7
19	2014		D01FK4130117	30	8	1993	2	20	4
20	2014		D01FK4130106	28	4	1993	1	20	8

Figure 4.3: Dataset in xlsx format

Data Preprocessing

Data preprocessing is a crucial step to ensure the dataset is clean, relevant, and suitable for analysis. The preprocessing steps undertaken include:

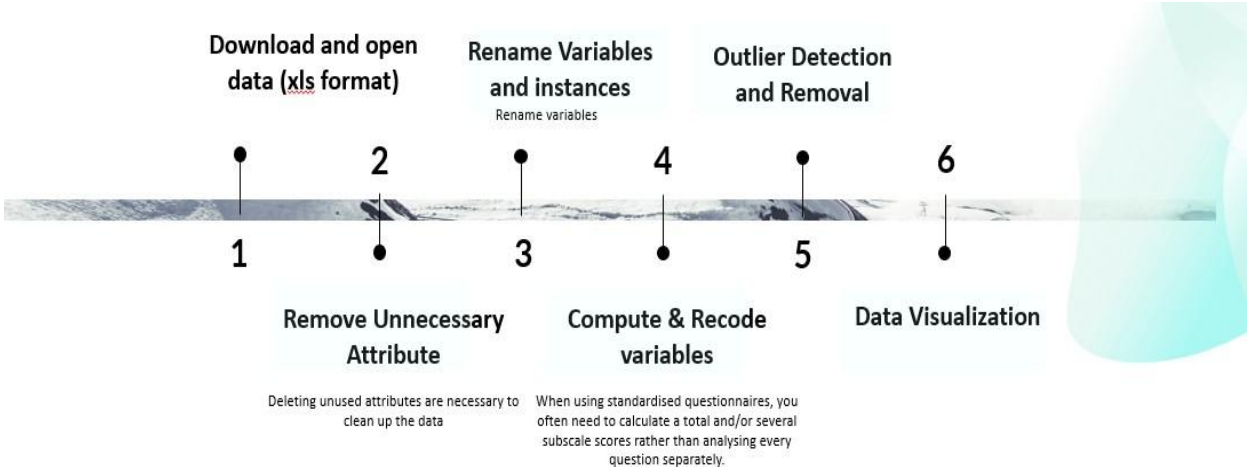


Figure 4.4: Steps in preprocessing

The selected attributes in Table 4.3 for the Community College Graduate Employability Dashboard Development were carefully chosen to provide a comprehensive analysis of the factors influencing graduate employability. These attributes cover critical areas such as demographics, academic performance, extracurricular involvement, curriculum relevance, career readiness, skill development, and employment status. They form the foundation for developing a data-driven dashboard to analyse and visualize employability outcomes effectively.

The dataset includes attributes related to the demographic and academic background of graduates, such as matriculation number, gender, age, study level, degree awarded, field of study, CGPA, university code, and graduation year. These attributes help to understand the educational and demographic diversity of community college graduates. Additionally, the dataset captures extracurricular involvement through attributes that measure participation in associations, clubs, and sports, offering insight into the role of co-curricular activities in improving employability.

The dataset also evaluates the relevance of academic programs by including attributes related to curriculum content, the balance between theory and practice, industrial training, and mandatory university courses. These factors help assess whether the community college curriculum effectively prepares graduates for the workforce. Furthermore, attributes related to career guidance and employment support are included to measure the availability and effectiveness of career counselling, job application assistance, interview preparation, and on-campus job opportunities. These insights are crucial for understanding the support systems that enhance employability.

Key skills and competencies are measured through attributes related to ICT proficiency, language skills (Malay and English), communication, critical thinking, problem-solving, teamwork, and general knowledge. These attributes are critical in assessing how well graduates are equipped with the necessary skills for professional environments. Finally, employment-related attributes such as job classification, workplace location, monthly income, employment sector, and reasons for unemployment provide valuable insights into employment outcomes for community college graduates.

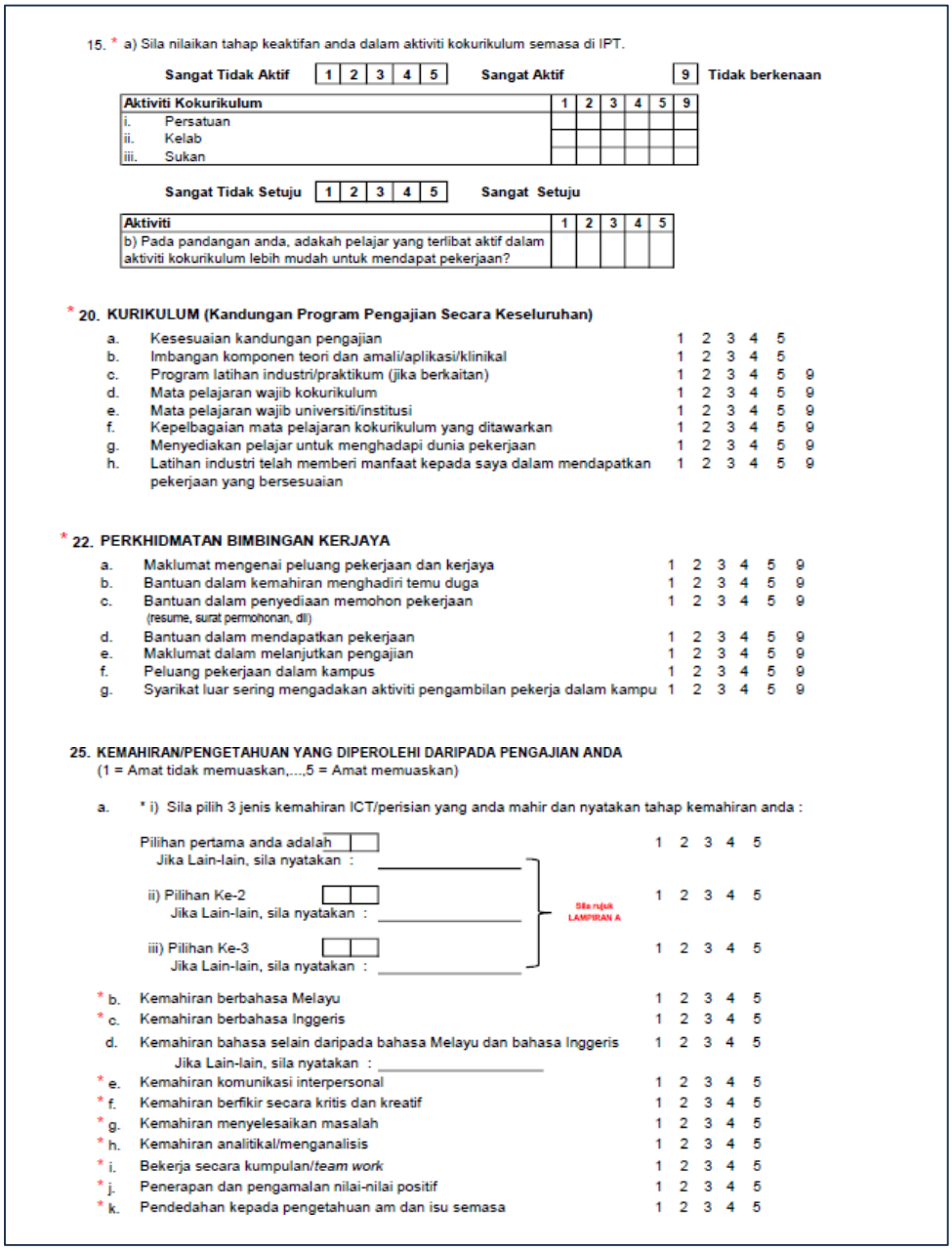
In conclusion, these attributes were selected to identify the key factors influencing employability, analyse the impact of education, skills, and support systems, and provide a clear and comprehensive view of graduate employability. By incorporating these attributes into the development of the dashboard, the project aims to create a tool that can help community colleges and policymakers understand and improve the employability of their graduates.

Table 4.3: Attributes Selection

No.	**Attributes**	**Descriptions**
1	e_matrik	Matriks
2	e_jantina	Jantina
3	e_umur	Umur
4	e_peringkat	Peringkat Pengajian
5	e_anugerah	Ijazah Yang Dianugerahkan
6	e_bidang	e_bidang
7	e_cgpa	CGPA / PNGK / HPNM / HPNG
8	e_kampus	Kod IPT
9	e_tahun	Tahun Konvokesyen
10	e_15_a_i	Penglibatan aktiviti ko-kurikulum (Persatuan)
11	e_15_a_ii	Penglibatan aktiviti ko-kurikulum (Kelab)
12	e_15_a_iii	Penglibatan aktiviti ko-kurikulum (Sukan)
13	e_15_b	Pandangan-aktif ko-kurikulum lebih mudah mendapat pekerjaan
14	e_20_a	Kesesuaian kandungan pengajian
15	e_20_b	Imbangan komponen teori dan amalan/aplikasi
16	e_20_c	Program latihan industri/praktikum (jika berkaitan)

17	e_20_d	Mata pelajaran wajib ko-kurikulum
18	e_20_e	Mata pelajaran wajib universiti/institusi
19	e_20_f	Kepelbagaian matapelajaran ko-kurikulum yang ditawarkan
20	e_20_g	Menyediakan pelajar untuk menghadapi dunia pekerjaan

Table 4.4 shows the selected attributes based on the questionnaire (Figure 4.5), which represent crucial aspects of a graduate's workforce readiness, including extracurricular participation, academic foundation, career guidance, and key skills. The new attributes Kokurikulum, Kurikulum, Bimbingan Kerjaya, and Kemahiran reflect important factors identified through the questionnaire that enhance employability.



15. * a) Sila nilaikan tahap keaktifan anda dalam aktiviti kokurikulum semasa di IPT.

Sangat Tidak Aktif 1 2 3 4 5 Sangat Aktif 9 Tidak berkenaan

Aktiviti Kokurikulum	1	2	3	4	5	9
i. Persatuan						
ii. Kelab						
iii. Sukan						

Sangat Tidak Setuju 1 2 3 4 5 Sangat Setuju

Aktiviti	1	2	3	4	5
b) Pada pandangan anda, adakah pelajar yang terlibat aktif dalam aktiviti kokurikulum lebih mudah untuk mendapat pekerjaan?					

* 20. KURIKULUM (Kandungan Program Pengajian Secara Keseluruhan)

a.	Kesesuaian kandungan pengajian	1	2	3	4	5	
b.	Imbangan komponen teori dan amali/aplikasi/klinikal	1	2	3	4	5	
c.	Program latihan industri/praktikum (jika berkaitan)	1	2	3	4	5	9
d.	Mata pelajaran wajib kokurikulum	1	2	3	4	5	9
e.	Mata pelajaran wajib universiti/institusi	1	2	3	4	5	9
f.	Kepelbagaian mata pelajaran kokurikulum yang ditawarkan	1	2	3	4	5	9
g.	Menyediakan pelajar untuk menghadapi dunia pekerjaan	1	2	3	4	5	9
h.	Latihan industri telah memberi manfaat kepada saya dalam mendapatkan pekerjaan yang bersesuaian	1	2	3	4	5	9

* 22. PERKHIDMATAN BIMBINGAN KERJAYA

a.	Maklumat mengenai peluang pekerjaan dan kerjaya	1	2	3	4	5	9
b.	Bantuan dalam kemahiran menghadiri temu duga	1	2	3	4	5	9
c.	Bantuan dalam penyediaan memohon pekerjaan (resume, surat permohonan, dll)	1	2	3	4	5	9
d.	Bantuan dalam mendapatkan pekerjaan	1	2	3	4	5	9
e.	Maklumat dalam melanjutkan pengajian	1	2	3	4	5	9
f.	Peluang pekerjaan dalam kampus	1	2	3	4	5	9
g.	Syarikat luar sering mengadakan aktiviti pengambilan pekerja dalam kampu	1	2	3	4	5	9

25. KEMAHIRAN/PENGETAHUAN YANG DIPEROLEHI DARIPADA PENGAJIAN ANDA
(1 = Amat tidak memuaskan,...,5 = Amat memuaskan)

a. * i) Sila pilih 3 jenis kemahiran ICT/perisian yang anda mahir dan nyatakan tahap kemahiran anda :

Pilihan pertama anda adalah [ ] 1 2 3 4 5
Jika Lain-lain, sila nyatakan : _____

ii) Pilihan Ke-2 [ ] 1 2 3 4 5
Jika Lain-lain, sila nyatakan : _____

iii) Pilihan Ke-3 [ ] 1 2 3 4 5
Jika Lain-lain, sila nyatakan : _____

Sila rujuk LAMPIRAN A

* b.	Kemahiran berbahasa Melayu	1	2	3	4	5
* c.	Kemahiran berbahasa Inggeris	1	2	3	4	5
d.	Kemahiran bahasa selain daripada bahasa Melayu dan bahasa Inggeris	1	2	3	4	5
	Jika Lain-lain, sila nyatakan : _____					
* e.	Kemahiran komunikasi interpersonal	1	2	3	4	5
* f.	Kemahiran berfikir secara kritis dan kreatif	1	2	3	4	5
* g.	Kemahiran menyelesaikan masalah	1	2	3	4	5
* h.	Kemahiran analitikal/menganalisis	1	2	3	4	5
* i.	Bekerja secara kumpulan/*team work*	1	2	3	4	5
* j.	Penerapan dan pengamalan nilai-nilai positif	1	2	3	4	5
* k.	Pendedahan kepada pengetahuan am dan isu semasa	1	2	3	4	5

Figure 4.5: Questionnaire from SKPG that relate with the

Table 4.4: Optimizing Attributes and Renaming

Attributes	New Attribute	Descriptions
e_15_a_i	kokurikulum	Penglibatan aktiviti ko-kurikulum (Persatuan)
e_15_a_ii		Penglibatan aktiviti ko-kurikulum (Kelab)
e_15_a_iii		Penglibatan aktiviti ko-kurikulum (Sukan)
e_15_b		Pandangan-aktif ko-kurikulum lebih mudah mendapat pekerjaan
e_20_a		Kesesuaian kandungan pengajian
e_20_b		Imbangan komponen teori dan amalan/aplikasi
e_20_c		Program latihan industri/praktikum (jika berkaitan)
e_20_d		Mata pelajaran wajib ko-kurikulum

e_20_e	kurikulum	Mata pelajaran wajib universiti/institusi
e_20_f		Kepelbagaian matapelajaran ko-kurikulum yang ditawarkan
e_20_g		Menyediakan pelajar untuk menghadapi dunia pekerjaan
e_20_h		Latihan Industri telah memberi manfaat kepada saya dalam mendapatkan pekerjaan bersesuaian
e_22_a	bimbingan_kerjaya	Maklumat mengenai peluang pekerjaan & kerjaya
e_22_b		Bantuan dalam kemahiran menghadiri temuduga
e_22_c		Bantuan dalam penyediaan memohon pekerjaan
e_22_d		Bantuan dalam mendapatkan pekerjaan
e_22_e		Bantuan dalam melanjutkan pengajian
e_22_f		Peluang pekerjaan dalam kampus
e_22_g		Proses pengambilan pekerja dalam kampus oleh majikan
e_25_a_1	kemahiran	Kod Kemahiran ICT1
e_25_b		Kemahiran berbahasa Melayu
e_25_c		Kemahiran berbahasa Inggeris
e_25_e		Kemahiran komunikasi interpersonal
e_25_f		Kemahiran berfikir secara kritis dan kreatif
e_25_g		Kemahiran menyelesaikan masalah
e_25_h		Kemahiran analitikal / menganalisis
e_25_i		Bekerja secara kumpulan / team work
e_25_j		Penerapan dan pengalaman nilai-nilai positif
e_25_k		Pendedahan kepada pengetahuan am dan isu semasa

The provided code as shown in Figure 4.6 demonstrates how to recode specific values in the e_peringkat column of a dataset using a predefined mapping dictionary. First, a dictionary called peringkat_mapping is created, which maps the values 7 and 71 in the e_peringkat column to more descriptive labels: 7 is recoded as "Sijil" and 71 as "Sijil Kolej Komuniti Bermodular (SKK(M))". Then, the .map() function is used to apply this mapping to the e_peringkat column of the dataset. The function replaces each value in the column with its corresponding mapped value from the dictionary, and if a value doesn't have a matching entry in the dictionary, it is set to NaN. Finally, the code prints the first five rows of the updated dataset to verify that the recoding has been successfully applied. This process helps in transforming numeric or coded values into more readable and meaningful labels, making the dataset easier to interpret and analyse.

Figure 4.6: Recode instances e_peringkat

```
# Mapping for e_peringkat
peringkat_mapping = {
    7: "Sijil",
    71: "Sijil Kolej Komuniti Bermodular (SKK(M))",
}

# recode
new_data.loc[:, 'e_peringkat'] = new_data.loc[:, 'e_peringkat'].map(peringkat_ma

# Verify the result
print(new_data.head(5))
```

Figure 4.7 is a `e_jantina` data preparation for effective visualization in Power BI. It is often necessary to convert numerical values that represent categories into their corresponding categorical labels. This process, known as encoding categorical variables, enhances readability and interpretability in visualizations, ensuring the data is presented in a more intuitive format for users.

```
# Mapping for e_jantina
jantina_mapping = {
    1: "Lelaki",
    2: "Perempuan"
}

# Create the new attribute jantina_mapped by mapping e_jantina
new_data.loc[:, 'e_jantina'] = new_data.loc[:, 'e_jantina'].map(jantina_mapping)


# Verify the result
print(new_data.head())
```

```
       e_matrik  e_jantina  e_umur                                e_peringkat  \
0   P03SKKM0003     Lelaki      21  Sijil Kolej Komuniti Bermodular (SKK(M))
1  D11FP5130105  Perempuan      20  Sijil Kolej Komuniti Bermodular (SKK(M))
2    R01LS35106  Perempuan      23  Sijil Kolej Komuniti Bermodular (SKK(M))
3    R01RK35039     Lelaki      20  Sijil Kolej Komuniti Bermodular (SKK(M))
4  P31PT6130119  Perempuan      20  Sijil Kolej Komuniti Bermodular (SKK(M))
```

Figure 4.7: Recode instances e_jantina

Recoding `e_umur` in Figure 4.8 involves grouping continuous age values into categorical bins (e.g., `18-24`, `25-34`) to simplify the data, making it easier to analyze and visualize. This transformation improves interpretability by providing a more intuitive representation of age groups, allowing for clearer comparisons and insights. It also facilitates better visualizations in tools like Power BI, where categorical data is easier to handle than raw numerical values. Additionally, recoding aligns with statistical analysis methods and demographic studies, which often rely on grouped data to identify trends across different age ranges, thus enhancing the overall effectiveness of the analysis.

```
# recode umur

lowest_age = new_data['e_umur'].min()
highest_age = new_data['e_umur'].max()

# Display the results
print(f"The lowest age is {lowest_age}.")
print(f"The highest age is {highest_age}.")


# Define the bin edges
bins = [18, 25, 35, 45, 55]  # Adjust these edges as needed
labels = ['18-24', '25-34', '35-44', '45-53']  # Labels for the bins

# Create a new column with age bins
new_data.loc[:, 'e_umur'] = pd.cut(new_data.loc[:, 'e_umur'], bins=bins, labels=

# Display the DataFrame
print(new_data)
```

```
The lowest age is 18.
The highest age is 53.
         e_matrik  e_jantina e_umur  \
0     P03SKKM0003     Lelaki  18-24
1    D11FP5130105  Perempuan  18-24
2      R01LS35106  Perempuan  18-24
3      R01RK35039     Lelaki  18-24
4    P31PT6130119  Perempuan  18-24
```

Figure 4.8: Recode instances e_umur

The attribute e_bidang was recoded to group academic disciplines into five primary categories Sastera & Sains Sosial, Sains, Teknikal, Teknologi Maklumat & Komunikasi, and Pendidikan as visualized in Figure 4.9. This recoding simplifies analysis by providing a structured classification of academic fields. As visualized in

```
# Mapping for e_bidang
bidang_mapping = {
    1: "Sastera & Sains Sosial",
    2: "Sains",
    3: "Teknikal",
    4: "Teknologi Maklumat & Komunikasi",
    5: "Pendidikan"
}

# recode
new_data.loc[:, 'e_bidang'] = new_data.loc[:, 'e_bidang'].map(bidang_mapping)

# Verify the result
print(new_data.head())
```

Figure 4.9: Recode instances e_bidang

The attribute Kumpulan Pekerjaan Utama in Figure 4.10 was recoded to classify job roles into nine broad categories: Pengurus, Profesional, Juruteknik dan Profesional Berkaitan, Pekerja Perkeranian Sokongan, Pekerja Perkhidmatan dan Jualan, Pekerja Mahir bidang Pertanian dan Perikanan, Pekerja Kraftangan dan Perdagangan Berkaitan, Operator Kilang dan Mesin dan Juru Pasang, and Pekerjaan Asas. This standardization enhances the interpretability of the data.

```
# recode for e_41_a kumpulan pekerjaan utama
e_41_a_mapping = {
    1: "Pengurus",
    2: "Profesional",
    3: "Juruteknik dan Profesional Berkaitan",
    4: "Pekerja Perkeranian Sokongan",
    5: "Pekerja Perkhidmatan dan Jualan",
    6: "Pekerja Mahir bidang Pertanian dan Perikanan",
    7: "Pekerja Kraftangan dan Perdagangan Berkaitan",
    8: "Operator Kilang dan Mesin dan Juru Pasang",
    9: "Pekerjaan Asas"
}

# recode
new_data.loc[:, 'e_41_a'] = new_data.loc[:, 'e_41_a'].map(e_41_a_mapping)

# Verify the result
print(new_data.head())
```

Figure 4.10: Recode instances e_41_a

The attribute Pendapatan in Figure 4.11 was recoded into eight salary ranges: RM500 dan ke bawah, RM501 - RM1000, RM1001 - RM1500, RM1501 - RM2000, RM2001 - RM2500, RM2501 - RM3000, RM3001 - RM5000, and RM5001. This recoding provides a clearer view of income levels and simplifies comparisons.

```
# recode for e_44 pendapatan
e_44_mapping = {
    1: "RM500 dan ke bawah",
    2: "RM501 - RM1000",
    3: "RM1001 - RM1500",
    4: "RM1501 - RM2000",
    5: "RM2001 - RM2500",
    6: "RM2501 - RM3000",
    7: "RM3001 - RM5000",
    8: "RM5001 dan ke atas"
}

# recode
new_data.loc[:, 'e_44'] = new_data.loc[:, 'e_44'].map(e_44_mapping)

# Verify the result
print(new_data.head())
```

Figure 4.11: Recode instances e_44

The attribute Sektor Pekerjaan in Figure 4.12 was recoded into broader categories, including Kerajaan, Badan Berkanun, Swasta Multinasional, Swasta Tempatan, Perusahaan Sendiri, Syarikat Berkaitan Kerajaan (GLC), Pertubuhan Bukan Kerajaan (NGO), and Lain-lain. This classification reduces redundancy and improves data interpretation

```
# Mapping for e_45 Sektor Pekerjaan

e_45_mapping = {
    1: "Kerajaan",
    2: "Badan Berkanun",
    3: "Swasta Multinasional",
    4: "Swasta Tempatan",
    5: "Perusahaan Sendiri",
    7: "Syarikat Berkaitan Kerajaan (GLC)",
    8: "Pertubuhan Bukan Kerajaan (NGO)",
    6: "Lain-lain"
}

# recode
new_data.loc[:, 'e_45'] = new_data.loc[:, 'e_45'].map(e_45_mapping)

# Verify the result
print(new_data.head())
```

Figure 4.12: Recode instances e_45

```
# Mapping for e_46_kod Sektor Ekonomi

e_46_kod_mapping = {
    1: "Pertanian, Perhutanan dan Perikanan",
    2: "Perlombongan dan Pengkuarian",
    3: "Pembuatan",
    4: "Bekalan elektrik, gas, wap dan pendingin udara",
    5: "Bekalan air; pembentungan, pengurusan sisa dan aktiviti pemulihan",
    6: "Pembinaan",
    7: "Perdagangan borong dan runcit, pembaikan kenderaan bermotor dan motosika
    8: "Pengangkutan dan Penyimpanan",
    9: "Penginapan dan aktiviti perkhidmatan makanan dan minuman",
    10: "Maklumat dan Komunikasi",
    11: "Aktiviti kewangan dan insurans/takaful",
    12: "Aktiviti hartanah",
    13: "Aktiviti profesional, saintifik dan teknikal",
    14: "Aktiviti pentadbiran dan khidmat sokongan",
    15: "Pentadbiran awam dan pertahanan; keselamatan sosial wajib",
    16: "Pendidikan",
    17: "Aktiviti kesihatan kemanusiaan dan kerja sosial",
    18: "Kesenian, hiburan dan rekreasi",
    19: "Aktiviti perkhidmatan lain",
    20: "Aktiviti isi rumah sebagai majikan bagi personel domestik; aktiviti men
    21: "Aktiviti badan dan pertubuhan luar wilayah"
}

# recode
new_data.loc[:, 'e_46_kod'] = new_data.loc[:, 'e_46_kod'].map(e_46_kod_mapping)

# Verify the result
print(new_data['e_46_kod'].head())
```

Figure 4.13: Recode instances e_46_kod

The attribute Sektor Ekonomi in Figure 4.13 was recoded to consolidate economic activities into 21 categories, such as Pertanian, Perhutanan dan Perikanan, Pembuatan, Pembinaan, and Aktiviti Kewangan dan Insurans/Takaful. This recoding organizes the dataset for easier analysis.

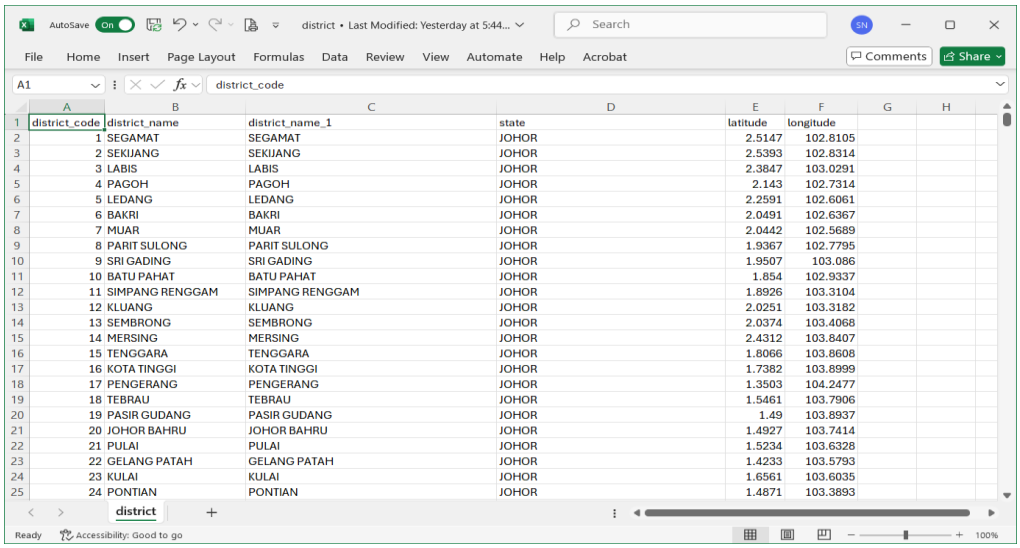
The attribute Status Kebolehpasaran in Figure 4.14 was recoded into five categories: Bekerja, Melanjutkan Pengajian, Meningkatkan Kemahiran, Menunggu Penempatan Pekerjaan, and Belum Bekerja. This recoding focuses on the core employability outcomes of graduates.

```
# Mapping for e_status to status_GE
status_mapping = {
    1: "Bekerja",
    2: "Melanjutkan Pengajian",
    3: "Meningkatkan Kemahiran",
    4: "Menunggu Penempatan Pekerjaan",
    5: "Belum Bekerja"
}

# Create a new column status_GE by mapping e_status
new_data['status_GE'] = new_data['e_status'].map(status_mapping).copy()

# Verify the result
print(new_data[['e_status', 'status_GE']].head())
```

```
   e_status status_GE
0         1   Bekerja
1         1   Bekerja
2         1   Bekerja
3         1   Bekerja
4         1   Bekerja
```

Figure 4.14: Recode instances e_status dan compute new attribute status_ge

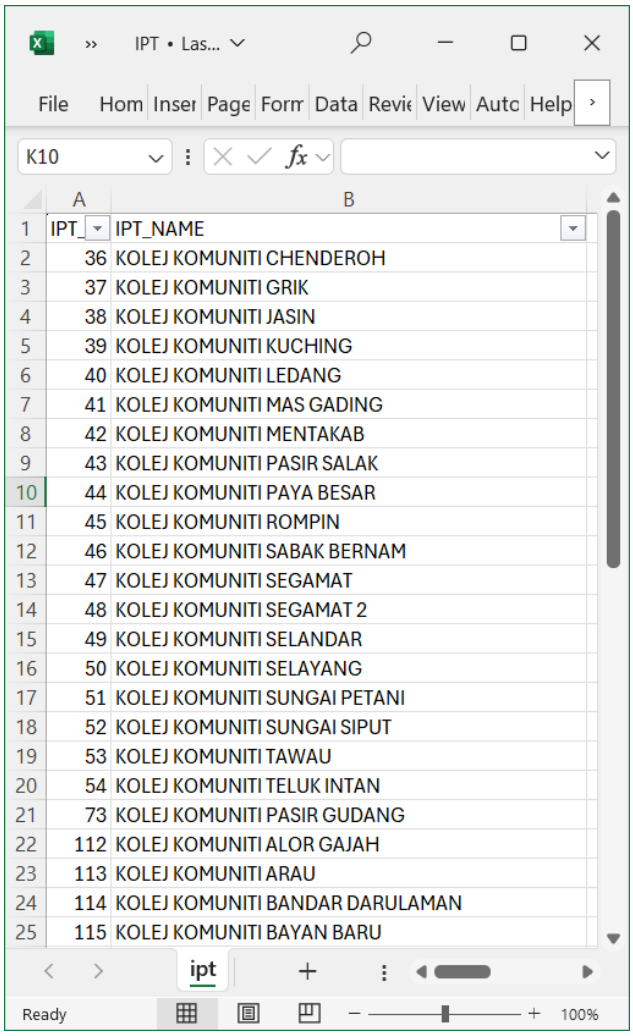
The district table in Figure 4.15 contains geographic information for each district, including a unique district_code, the district's name in two formats (district_name and district_name_1), the state it is in, and its geographic coordinates (latitude and longitude). This table is crucial for mapping and analyzing locations, as it allows for precise identification of districts and enables spatial analysis by using the latitude and longitude data. Additionally, it can be used to filter data based on specific districts or regions.



district_code	district_name	district_name_1	state	latitude	longitude
1	SEGAMAT	SEGAMAT	JOHOR	2.5147	102.8105
2	SEKIJANG	SEKIJANG	JOHOR	2.5393	102.8314
3	LABIS	LABIS	JOHOR	2.3847	103.0291
4	PAGOH	PAGOH	JOHOR	2.143	102.7314
5	LEDANG	LEDANG	JOHOR	2.2591	102.6061
6	BAKRI	BAKRI	JOHOR	2.0491	102.6367
7	MUAR	MUAR	JOHOR	2.0442	102.5689
8	PARIT SULONG	PARIT SULONG	JOHOR	1.9367	102.7795
9	SRI GADING	SRI GADING	JOHOR	1.9507	103.086
10	BATU PAHAT	BATU PAHAT	JOHOR	1.854	102.9337
11	SIMPANG RENGGAM	SIMPANG RENGGAM	JOHOR	1.8926	103.3104
12	KLUANG	KLUANG	JOHOR	2.0251	103.3182
13	SEMBRONG	SEMBRONG	JOHOR	2.0374	103.4068
14	MERSING	MERSING	JOHOR	2.4312	103.8407
15	TENGGARA	TENGGARA	JOHOR	1.8066	103.8608
16	KOTA TINGGI	KOTA TINGGI	JOHOR	1.7382	103.8999
17	PENGERANG	PENGERANG	JOHOR	1.3503	104.2477
18	TEBRAU	TEBRAU	JOHOR	1.5461	103.7906
19	PASIR GUDANG	PASIR GUDANG	JOHOR	1.49	103.8937
20	JOHOR BAHRU	JOHOR BAHRU	JOHOR	1.4927	103.7414
21	PULAI	PULAI	JOHOR	1.5234	103.6328
22	GELANG PATAH	GELANG PATAH	JOHOR	1.4233	103.5793
23	KULAI	KULAI	JOHOR	1.6561	103.6035
24	PONTIAN	PONTIAN	JOHOR	1.4871	103.3893

Figure 4.15: Latitude and Longitude Computation for Dataset district

The IPT dataset in Figure 4.16 is provided for mapping with the codes in the main dataset. This table contains two columns: IPT_ID and IPT_NAME. The data is required for filtering the dataset by institution in dashboard visualization.



IPT_	IPT_NAME
36	KOLEJ KOMUNITI CHENDEROH
37	KOLEJ KOMUNITI GRIK
38	KOLEJ KOMUNITI JASIN
39	KOLEJ KOMUNITI KUCHING
40	KOLEJ KOMUNITI LEDANG
41	KOLEJ KOMUNITI MAS GADING
42	KOLEJ KOMUNITI MENTAKAB
43	KOLEJ KOMUNITI PASIR SALAK
44	KOLEJ KOMUNITI PAYA BESAR
45	KOLEJ KOMUNITI ROMPIN
46	KOLEJ KOMUNITI SABAK BERNAM
47	KOLEJ KOMUNITI SEGAMAT
48	KOLEJ KOMUNITI SEGAMAT 2
49	KOLEJ KOMUNITI SELANDAR
50	KOLEJ KOMUNITI SELAYANG
51	KOLEJ KOMUNITI SUNGAI PETANI
52	KOLEJ KOMUNITI SUNGAI SIPUT
53	KOLEJ KOMUNITI TAWAU
54	KOLEJ KOMUNITI TELUK INTAN
73	KOLEJ KOMUNITI PASIR GUDANG
112	KOLEJ KOMUNITI ALOR GAJAH
113	KOLEJ KOMUNITI ARAU
114	KOLEJ KOMUNITI BANDAR DARULAMAN
115	KOLEJ KOMUNITI BAYAN BARU

Figure 4.16: Dataset IPT

Descriptive Statistics and Data Visualization

Summary statistics were generated to understand the distribution of each attribute, providing key insights into the central tendencies, variability, and overall shape of the data. For instance, measures such as the mean, median, mode, range, variance, and standard deviation were calculated to offer a comprehensive overview of each attribute's distribution. The dataset includes 23 attributes, such as gender, age, program, CGPA, and employability status, each contributing valuable information for analysis.

To complement the summary statistics, various plots were created to visualize the distribution and relationships between attributes. Histograms were used to display the frequency distribution of continuous variables like age and CGPA, highlighting patterns such as skewness and the presence of outliers.

Scatter plots were employed to examine potential correlations between pairs of numerical attributes, such as CGPA and employability status, revealing trends and relationships that might not be evident from summary statistics alone. Additionally, bar charts were utilized to represent categorical data, such as gender and employability status, making it easier to compare frequencies and distributions across different categories.

These visualizations, combined with the summary statistics, facilitated a deeper understanding of the data's structure and relationships, aiding in the identification of key patterns and insights that would inform further analysis and decision-making.

The data on Figure 4.17 reveals the following distribution of employment status: A significant majority of graduates, 2,328, are employed. In contrast, 586 graduates are continuing their studies. Only 182 graduates are not employed. This distribution suggests a strong employability rate among the graduates, with a large proportion actively engaged in the workforce. The relatively smaller number of graduates not in employment highlights the effectiveness of the current educational and career preparation systems in facilitating job placement or further academic pursuits.

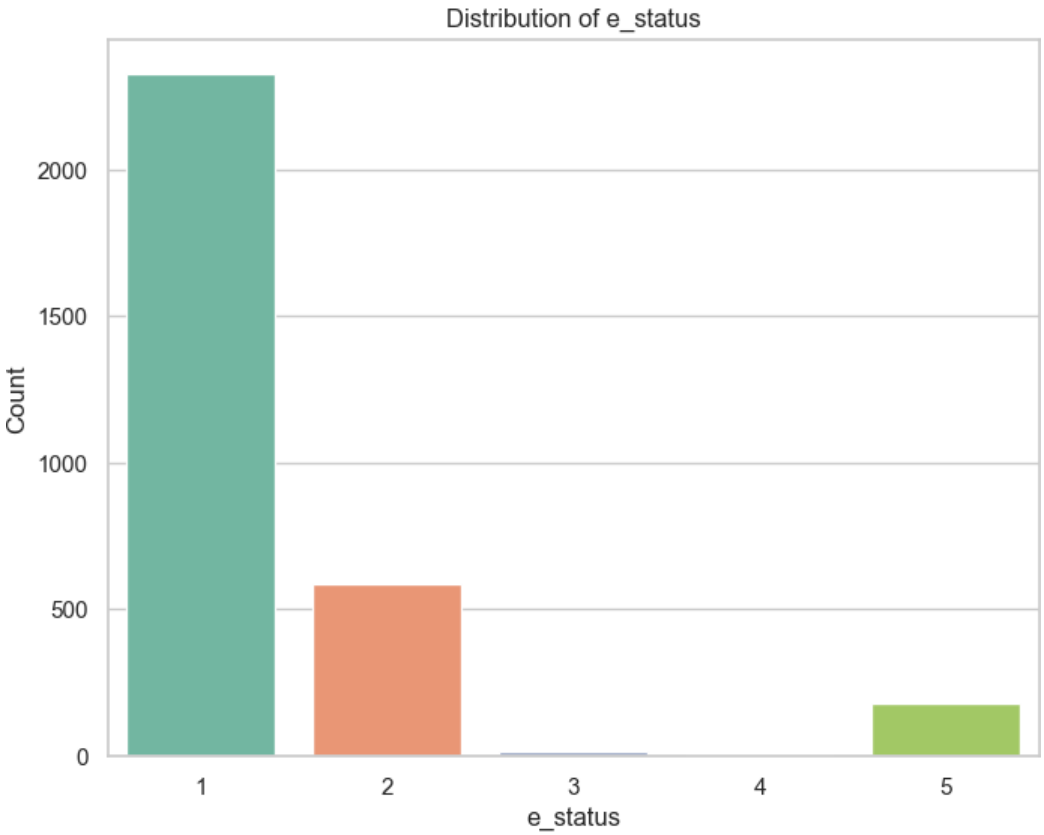


Figure 4.17: EDA of employment status

The histogram in Figure 4.18 provides a visual representation of the distribution of CGPA values among the datasets. The histogram shows that CGPA values range from a minimum of 2.0 to a maximum of 4.0. This indicates that all data points fall within this range, and there are no outliers beyond these bounds. The average CGPA is 3.4. This suggests that, on average, students have a relatively high CGPA, which could imply a strong overall academic performance in the dataset.

The histogram shows a left-skewed distribution of CGPA values, meaning most students have high CGPAs, with fewer students having lower CGPAs. This indicates that most students perform well academically, and lower CGPAs are less common in this dataset.

The histogram reveals that most students have high CGPA values, indicating strong academic performance. This positive trend suggests that students are generally performing well, which is encouraging for the educational institution or program.

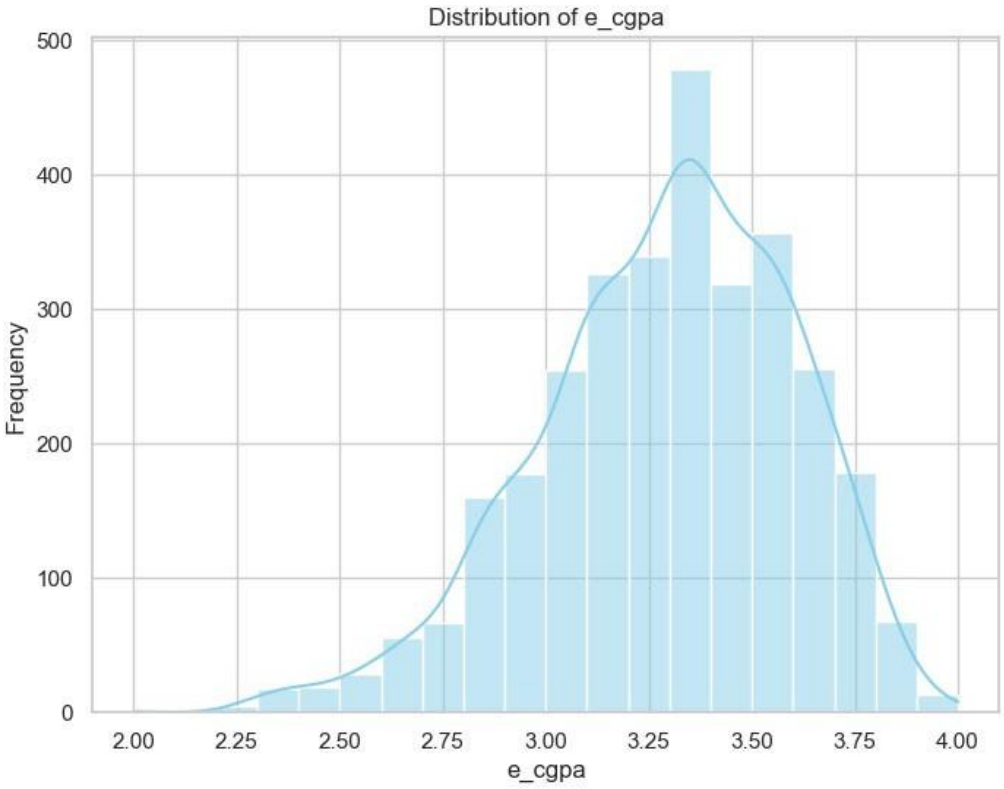


Figure 4.18: Histogram of CGPA distribution

The gender distribution among graduates in Figure 4.19 shows a relatively balanced employment status, with 1,887 male graduates and 1,231 female graduates. This indicates a fairly even distribution between genders in terms of employment outcomes.

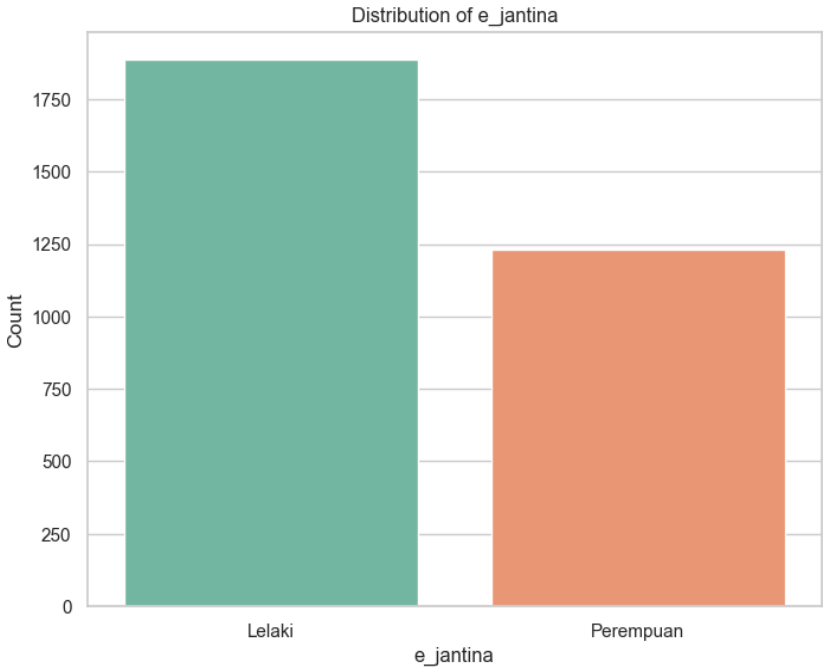


Figure 4.19 EDA of gender distribution

Graduates from certificate programs and SKMM (Sijil Kemahiran Malaysia) exhibited higher employment rates, with 13 graduates from certificate programs and 3,105 from SKMM being employed as shown in Figure 4.20.

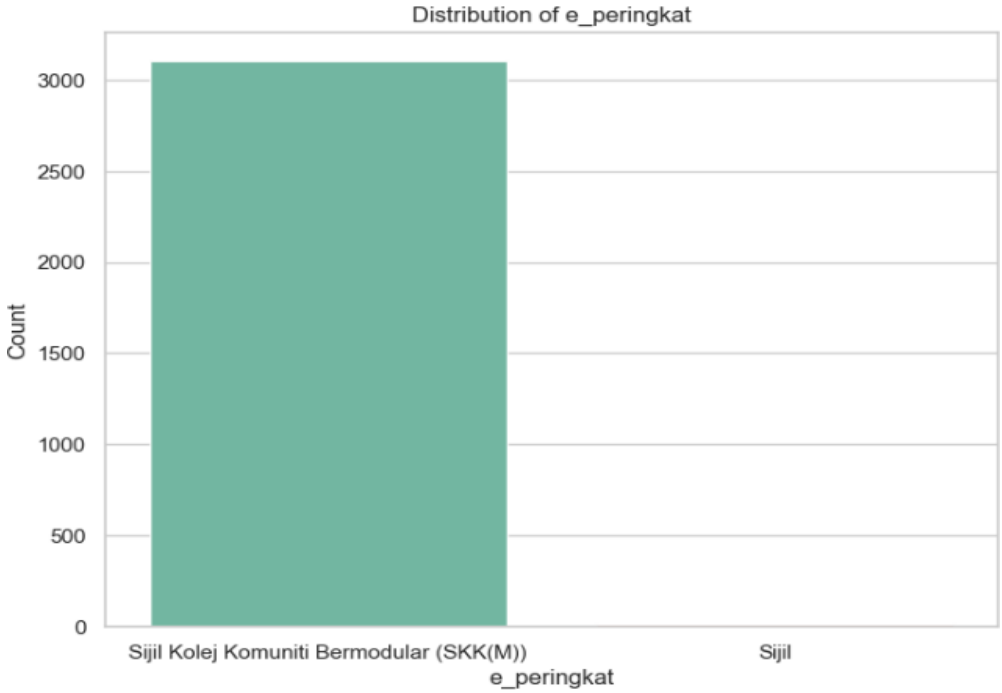


Figure 4.20 Data distribution by Program Levels: Graduates

Based on the Figure 4.21, the distribution of job sectors is as follows: The Swasta Tempatan (Local Private Sector) category has the highest number of instances with 1,438 entries. This is significantly higher compared to other categories. Following this, Perusahaan Sendiri (Own Business) has 411 entries, while Swasta Multinasional (Multinational Private Sector) has 265. The Kerajaan (Government) sector has 60 entries, and NGO (Non-Governmental Organizations) has 45 entries. Badan Berkanun (Statutory Bodies) has the smallest number with 23 entries. Other categories include Lain-lain (Others) with 55 entries and GLC (Government-Linked Companies) with 31 entries.

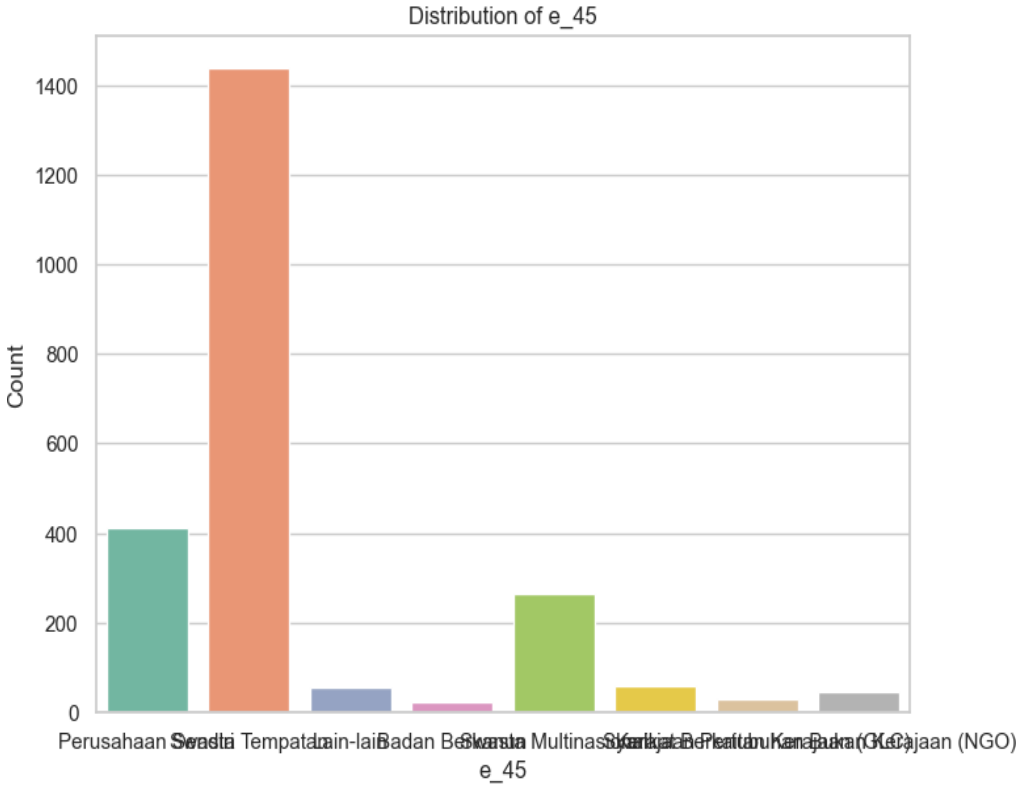


Figure 4.21: EDA of job sector

Handling Missing Values and Outliers

Missing values were addressed through imputation or removal to ensure data integrity. In this dataset, there is missing values were found, as shown in the Figure

4.22. The attributes with missing values were not replaced with any values because, the missing data was found to be Not Applicable. Additionally, these attributes are not used for calculations but are solely intended for descriptive analysis.

```
# Check for missing values

# Filter columns with missing values
missing_columns = new_data.columns[new_data.isnull().any()]

# Display attributes with missing values and their count
missing_summary = new_data[missing_columns].isnull().sum()

print("Attributes with missing values:")
print(missing_summary)
```

```
Attributes with missing values:
e_41_a           790
e_41e_daerah     790
e_44             790
e_45             790
e_46_kod         790
e_54            2328
dtype: int64
```

Figure 4.22: Checking missing values

The provided code as shown in the Figure 4.23 defines a function called handle_outliers, which is used to remove outliers from specified numerical columns in a dataset using the Interquartile Range (IQR) method. First, the function calculates the first quartile (Q1) and third quartile (Q3) of the data for a given column and then computes the IQR by subtracting Q1 from Q3. Using the IQR, the lower and upper bounds for identifying outliers are determined by subtracting 1.5 times the IQR from Q1 for the lower bound and adding 1.5 times the IQR to Q3 for the upper bound. The clip function is then applied to the column, which limits the values to fall within these bounds, effectively removing outliers by capping extreme values. The function is applied to a list of numerical columns such as e_cgpa, kokurikulum, kurikulum, bimbingan_kerjaya, and kemahiran. This approach helps in cleaning the data by addressing outliers, which could otherwise distort statistical analysis or model performance.

Figure 4.23: Running IQR using Python

```
import pandas as pd
import numpy as np

# Load your dataset
data = pd.read_csv('ge.csv')

# Select only numeric columns
numeric_data = data.select_dtypes(include=[np.number])

# Identify outliers using the IQR method
Q1 = numeric_data.quantile(0.25)
Q3 = numeric_data.quantile(0.75)
IQR = Q3 - Q1

# Filter out outliers
outliers = ((numeric_data < (Q1 - 1.5 * IQR)) | (numeric_data > (Q3 + 1.5 * IQR))).any(axis=1)
non_outliers_data = numeric_data[~outliers]

# Optionally, analyze the effect of removing outliers
print(f"Original numeric data shape: {numeric_data.shape}")
print(f"Numeric data shape without outliers: {non_outliers_data.shape}")

# Analyze the effect on key statistics
print(f"Mean with outliers: {numeric_data.mean()}")
print(f"Mean without outliers: {non_outliers_data.mean()}")
```

CONCLUSION

The primary objectives of this research were to perform data cleaning on the Graduate Tracer Study KPT dataset, develop and evaluate predictive models for forecasting graduate employability outcomes, and create a Power BI dashboard for data visualization and decision support. The research outcomes are as follows:

Successfully preprocessed and cleaned the Graduate Tracer Study dataset for the years 2014, ensuring the removal of inaccuracies, inconsistencies, and missing values. This process improved the dataset's quality, making it suitable for further analysis.

Developed and evaluated several predictive models to forecast graduate employability outcomes. The models demonstrated strong performance metrics, providing reliable predictions that can assist in understanding factors influencing employability.

Created a comprehensive Power BI dashboard that visualizes the cleaned data and model predictions. The dashboard includes various interactive elements, allowing stakeholders to explore employment trends, identify significant factors, and make informed decisions based on the insights provided. Bar charts, heatmaps, and boxplots were used to reveal trends and distributions within the dataset. These visualizations helped highlight critical aspects such as income levels, employment sectors, and economic activity distribution among graduates.

In summary, the EDA process provided a comprehensive understanding of the dataset, laying the groundwork for model development in subsequent chapters. By addressing data quality, exploring attribute relationships, and visualizing key patterns, this chapter established a solid foundation for predictive and diagnostic analytics in the study of graduate employability.

REFERENCES

1. Ab Rahman, M. S. (2023), Tiga Faktor Kebolehpasaran Graduan-Pascapandemik, Kosmo, retrieved from https://www.kosmo.com.my/2023/11/29/tiga-faktor-kebolehpasaran-graduan-pascapandemik/
2. ElSharkawy, G., Helmy, Y., and Yehia, E. (2022), Employability Prediction of Information Technology Graduates using Machine Learning Algorithms, International Journal of Advanced Computer Science and Applications, (IJACSA), Vol. 13, No. 10, 2022 359 | P a g e, www.ijacsa.thesai.org
3. Fuad, F. (2020), Kebolehpasaran Graduan TVET Capai 95 Peratus, Berita Harian, retrieved from https://www.bharian.com.my/berita/nasional/2020/11/759851/ kebolehpasaran-graduan-tvet-capai-95-peratus/
4. Haque, R., Quek, A., Ting, C. Y., Goh, H. N., & Hasan, M. R. (2024), Classification Techniques Using Machine Learning for Graduate Student Employability Predictions. International Journal on Advanced Science, Engineering & Information Technology, 14(1).
5. Ibrahim, A., (2023), Kebolehpasaran Graduan Satu Cabaran Nasional, Utusan Malaysia, retrieved from https://www.utusan.com.my/nasional/2023/08/ kebolehpasaran-graduan-satu-cabaran-nasional/
6. Jabatan Penerangan Malaysia (2024), Gaji premium lepasan TVET ditetapkan hingga RM4,000, Jabatan Penerangan Malaysia, retrieved from https://www.penerangan.gov.my/gaji-premium-lepasan-tvet-ditetapkan-pada-kadar-rm2500-hingga-rm4000/
7. Jabarullah, N. H. and Iqbal Hussain, H. (2019), The Effectiveness of Problem-Based Learning in Technical and Vocational Education in Malaysia, Education + Training, Vol. 61 No. 5, pp. 552-567. https://doi.org/10.1108/ET-06-2018-0129
8. Shahriyar, J., Ahmad, J. B., Zakaria, N. H. and Su, G. E. (2022), Enhancing Prediction of Employability of Students: Automated Machine Learning Approach, 2022 2nd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA), Bandung, Indonesia, pp. 87-92, doi: 10.1109/ICICyTA57421.2022.10038231.
9. MDEC (Malaysia Digital Economy Corporation). (2021). Annual Report 2021.
10. Ministry of Higher Education Malaysia. (2021). Laporan Kajian Pengesanan Graduan 2021. Ministry of Higher Education Malaysia.
11. Ministry of Higher Education Malaysia. (2022). Laporan Kajian Pengesanan Graduan 2022. Ministry of Higher Education Malaysia.
12. Ministry of Higher Education Malaysia. (2023). Laporan Kajian Pengesanan Graduan 2023. Ministry of Higher Education Malaysia.
13. Ministry of Higher Education Malaysia. (2020). Pelan Strategik Kebolehpasaran Graduan KPT 2021-2025. Ministry of Higher Education Malaysia.

14. Ministry of Education Malaysia. (2019). Technical and Vocational Education and Training (TVET) Strategy 2018-2025. Ministry of Education Malaysia.
15. Raman,R. and Pramod, D. (2021), The role of predictive analytics to explain the employability of management graduates, Emerald Insight, https://www.emerald.com/insight/1463-5771.htm
16. Sapaat, M. A., Mustapha, A., Ahmad, J., Chamili, K. and Muhamad, R. (2011). A Data Mining Approach to Construct Graduates Employability Model in Malaysia. International Journal of New Computer Architectures and their Applications (IJNCAA). 4. 1111-1124.
17. Tableau Software. (2020). Best Practices for Creating Effective Dashboards. Available at: Tableau.
18. UNESCO. (2016). Strategy for Technical and Vocational Education and Training (TVET) 2016-2021. UNESCO.
19. Vinutha, K. and Yogisha, H. K. (2021), Prediction of Employability of Engineering Graduates using Machine Learning Techniques, 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, pp. 742-745.
20. Zheng D. (2023), Simulation Research on College Students' Employment Prediction Model Based on Decision Tree Classification Algorithm, 2023 International Conference on Internet of Things, Robotics and Distributed Computing (ICIRDC), Rio De Janeiro, Brazil, pp. 194-199, doi: 10.1109/ICIRDC62824.2023.00041.