Journal of Informatics and Web Engineering
Vol. 2 No. 2 (September 2023) eISSN: 2821-370X

Workplace Preference Analytics Among Graduates

Sin-Yin Ong*, Choo-Yee Ting, Hui-Ngo Goh, Albert Quek, Chin-Leei Cham
Faculty of Computing and Informatics, Multimedia University, Malaysia
**corresponding author: (1181203333@student.mmu.edu.my, ORCiD: 0009-0006-2051-6532)*

Abstract - Graduates often find themselves difficult to secure a job after completing their education at universities or colleges. In this light, researchers have proposed various solutions to address this challenge. However, most of the work has largely focused on academic profile and personality traits; very few have highlighted the importance of workplace location characteristics. To address this challenge, this study has employed feature selection and machine learning approach to help graduates identify desired company type and sector based on their preferences and preferred location. The data used in this study was obtained from the Ministry of Higher Education Graduates Tracer Study's data, specifically for 2382 Multimedia University (MMU) students' employment situation upon graduating. Additional analytical datasets focusing on company and graduate locations were developed in order to extract further features relevant for this analysis. Feature selection was used to identify top-10 predictors that influence the selection of jobs in graduates' desired sectors. Various analytics methods such as Decision Tree Analysis, Random Forest Model selection, Naive Bayes Classification Method, Support Vector Machines and K-Nearest Neighbor Algorithms were employed for comparative evaluations within the workplace analytics scope. Notably so, results from this study demonstrate that using Random Forest Algorithm resulted in better performance in predicting employment status with an accuracy rate of 99.40%, predicting company type with 66.60% and lastly predicting company sector with 30.80% when compared to other predictive models utilized during our research work's project lifecycle phase.

Keywords—classification, machine learning classifier, data visualization, feature selection, workplace preference analytics

Received: 16 June 2023; Accepted: 25 August 2023; Published: 16 September 2023

I. INTRODUCTION

Workplace is a place where a university or college graduate steps into after completing their studies. Finding a suitable workplace has always been a challenge for a fresh graduate [1] which ultimately contributes to unemployment. This is due to the lack of working experience and skill sets [2]. Search phenomenon can be seen from the report in the tracer studies conducted by the Ministry of Higher Education, the Economic Planning Unit of the Malaysian government, which indicated that 25% of fresh graduates in the country are still jobless six months after getting their degrees [3]. Therefore, finding a suitable workplace is crucial. Unemployment is a crucial issue that higher education institutions should pay attention to in order to help graduates prepare for employment [4]. In order to facilitate the process of graduates finding a suitable job, workplace preference analytics is proposed in this study to analyze graduates' preferences on the basis of their profiles, especially workplace location through machine learning approaches. This research concentrates on using their preferences and ideal geographic location to identify their desired company type and sector.

Journal of Informatics and Web Engineering
https://doi.org/10.33093/jiwe.2023.2.2.17

Published by MMU Press. URL: https://journals.mmupress.com/jiwe

Firstly, the first vital problem of this study is to identify the optimal factors for choosing a suitable job sector from numerous existing factors. Consider factors including conditions of the work environment, promotion opportunities, salary, organization culture, employee health and safety, and many more aspects [2]. In fact, recent graduates do not have a clear plan for their future careers. Most of them have the least understanding of the job market and haven't taken the time to thoroughly consider the factors [4].

Secondly, there is another challenge in locating reliable data points to construct effective predictive models for workplace preference prediction. This research is looking at their profile, conditions and specific geographic preferences all at once to find out where they would be most preferable to work at. However, most of the graduates don't have any work experience as is normally stated blank in their career history [5] and the profiles may not reflect a complete representation of the characteristics of a graduate [1], so there is limited data to be utilized. Therefore, it is important to determine which features can significantly impact the prediction to get a trustworthy result.

Thirdly, workplace preference analytics is built which intends to recommend an ideal workplace that is a good fit for graduates. This makes improved methods of workplace matching for graduates currently to be expected in high demand. Therefore, it is essential to identify which are the suitable predictive models that should be built for the purpose of suggesting recent graduates onto a career that best suits their interests. This is to tackle the issue of the process of the job application being slowed down due to the difficulty of sourcing skilled graduates that match the company's needs [6]. With a better method, the suitability of workplace selection could be enhanced accordingly.

In summary, there are three objectives in developing a workplace preference analytics system as follow:

a) To identify the optimal factors of selecting an appropriate job sector in helping graduates to make informed decisions.
b) To determine the reliable data points that make predictive models perform well in finding their desired workplace.
c) To find out which predictive models are suitable to be developed in suggesting graduates an ideal workplace based on graduates' profile obtained.

II. LITERATURE REVIEW

A. Reasons for Unemployment among Fresh Graduates

The need of constructing workplace preference analytics is caused by a variety of factors. The suitability of a workplace can be discovered from the factors of unemployment among fresh graduates. Thus, related studies by a number of researchers take into account the difficulties faced by graduates as underlying causes while developing the models to fit the graduates' needs on workplace selection, as shown in Table 1 below.

Table 1: Reasons for Unemployment among Fresh Graduates

References	Insufficient employment skills	Employment dissatisfaction	Employment data disorganised	Unclear employment planning	High competitiveness for a job opportunity	Time consuming on job application
[7] 2017	X			X		
[3] 2018			X			X
[8] 2018		X				
[9] 2018			X			
[5] 2019	X	X		X		X
[10] 2019					X	X

[11] 2019		X				
[12] 2019			X		X	
[13] 2019	X					
[14] 2019	X					
[15] 2019				X		
[16] 2019			X			
[17] 2019			X			
[1] 2020		X				
[18] 2020				X		
[19] 2020		X				
[20] 2020	X					
[21] 2020	X		X		X	
[2] 2021	X				X	
[4] 2021	X			X		
[6] 2021		X	X			
[22] 2021	X					
[23] 2021		X				
[24] 2021	X	X				
[25] 2021		X		X		
[26] 2022	X	X				X
Total count of each unemployment reason	**11**	**10**	**7**	**6**	**4**	**4**

Table 1 provides an overview of the 26 studies' findings on the reason fresh graduates remain unemployed. A trend that can be observed in the recent six years is the number of papers that mention insufficient employment skills and employment dissatisfaction is getting higher. Meanwhile, the issue of employment data being disorganized, graduates having unclear employment planning and high competitiveness for a job opportunity are less common in recent three years, but time-consuming job applications can be seen to cause employment issues in 2018 and 2019.

A primary problem raised by the majority of researchers is insufficient employment skills which are often struggled by unemployed graduates. This graduates' challenge has been highlighted in 11 out of 26 papers. Recent fresh graduates always find it hard to get employed due to their inexperience in the workforce, therefore they should seek job positions that closely align with their present qualifications [2]. Despite there being relevant courses regarding teaching practice in the talent development programmes, graduates still do not possess soft skills like management skills, or even emergency handling [4]. Accordingly, they frequently fail on getting employed due to their criteria falling short of what the industry demands.

Besides, the secondary focused reason that was highlighted by researchers is employment dissatisfaction in 10 out of 26 papers. Graduates often have trouble finding a satisfactory job that matches their expectations upon entering the workforce, which may lead to low productivity and life dissatisfaction. Many factors should be considered while choosing an appropriate job sector, including the working environment, starting wage, promotion opportunities, annual compensation increases, bonuses, and other perks [25]. However, companies also struggle to find recent graduates who possess the right mix of skill sets and personality traits to fill open positions [6]. Employment dissatisfaction has the potential to reduce the productivity of employees and companies, which may cause an impact on national economic growth.

Lastly, the tertiary factor concentrated is the disorganized employment data which appears in 7 out of the papers. Researchers found that even with the recent improvements in computer resources, the growth of data still causes problematic aspects in handling and analyzing it [16]. Employment data disorganization must be a serious issue for all organizations, notably for businesses and educational institutions. Although graduates fill up their employment status into an online system which is required by most universities, the data might not be updated and accurate since the process is time-consuming and they might not even remember [3]. Consequently, it is difficult to keep track of the employability of the graduates.

In summary, the literature review provided us with a clearer insight into the graduates' issues to be concentrated on. To solve the challenges that have been raised by several studies above, workplace preference analytics is proposed to provide a suggested workplace that matches the current criteria of the graduates, thereby assisting in addressing the

issue of their lack of work experience and skill sets as well as finding a potentially satisfying job for them. Lastly, employment data supplied by the graduates could be organized and arranged in a better structure for model construction and further analysis.

B. Machine Learning Techniques for Workplace Analytics

Machine Learning can be defined as a decision-support process which mainly leverages artificial intelligence, machine learning, statistics, pattern recognition, and database by extracting useful information from a large amount of data [19]. Businesses can utilize the models and hidden patterns discovered for risk assessment and forecasting as decision-makers to produce trustworthy market strategies such as minimizing production costs and improving their competitive advantage [9]. Several related research has been presented and conducted with numerous methods through machine learning approaches. A summary of related studies using machine learning techniques for workplace analytics is shown in Table 2 below:

Table 2: Machine Learning Techniques for Workplace Analytics

	Classification						Cluster -ing		Ensemble Learning			Deep Learning		
Author	Decision Tree (DT)	Random Forest (RF)	Naive Bayes (NB)	Support Vector Machines (SVM)	K-Nearest Neighbor (k-NN)	Logistic Regression (LR)	K-means	K-DBSCAN	Boosting	Stacking	Bagging	Multilayer perceptron (MLP)	Convolutional neural network (CNN)	Long Short-Term Memory (LSTM)
[7] 2017	X	X		X		X								
[27] 2017	X	X			X				X					
[9] 2018	X													
[28] 2018	X		X		X									
[29] 2018	X	X		X								X		
[30] 2018		X		X										
[31] 2018				X										
[5] 2019							X							
[11] 2019							X							
[17] 2019								X						
[1] 2020							X							
[19] 2020	X													
[32] 2020			X											
[33] 2020			X											
[34] 2020	X		X											
[35] 2020										X			X	X
[2] 2021							X							
[6] 2021	X	X	X	X		X								
[24] 2021	X		X							X	X	X		
[25] 2021							X							
[36] 2021	X													

[37] 2021	X	X	X	X	X				X					
[38] 2021		X												
[39] 2021														X
[26] 2022		X												
[40] 2022						X	X						X	
Total count of each technique	**11**	**8**	**7**	**6**	**3**	**3**	**6**	**1**	**2**	**2**	**1**	**2**	**2**	**2**

Table 2 indicates a summary of techniques employed for workplace analytics from 26 research. Classification algorithms are the most trending techniques in the recent six years, while clustering, ensemble learning, and deep learning are less proposed by researchers. Decision Tree, Random Forest, Naive Bayes and Support Vector Machine took place on the most proposed techniques for constructing workplace analytics.

First of all, the primary method proposed by the majority of researchers is classification algorithms. A classification algorithm is a supervised learning technique used in data mining which uses labellel input data to predict the likelihood or probability such that subsequent data will fit into the predetermined categories. The education institutions have widely made extensive use of classification strategies to classify their interests, abilities and behaviors [20]. In the literature review, there are 6 classification classifiers commonly proposed by the researchers including Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), Support Vector Machines (SVM), K-Nearest Neighbor (k-NN) and Logistic Regression (LR). According to [6], 5 classification models were built for predicting job employment status. When model parameters were optimized, Naive Bayes went from 88% accuracy to 91% accuracy while Random forest's accuracy rose from 91% to 93%, making it competitive with the support vector machine classification model which outperformed other classifiers with 93%.

Secondly, cluster analysis is the secondary technique that focuses on the aspect of the research that the researchers are least interested in, which is used for analyzing the location of their job selection. Cluster analysis is the process of dividing unlabelled data into distinct groups, or clusters, depending on the similarities between them [1]. According to the research conducted by [25], cluster analysis on the job sector was being implemented using K-Means clustering. Therefore, comprehensive data is being explored to gather similar potential job sectors. The traditional “Elbow Method” is used to find the optimal cluster number, thus this study concluded that 5 clusters would be sufficient to segregate the potential job sectors.

Thirdly, the least of the researchers have proposed other methods such as ensemble learning techniques and deep learning models. According to [35], ensemble methods have been constructed by combining multiple outputs of different deep neural network models so-called stacking to improve the performance of the prediction. The IT job prediction was implemented using the TextCNN model as well as more complex models like Bi-GRU-LSTM-CNN and Bi-GRU-CNN with the F1-score of 71.72%. In addition, [24] have developed an ensemble model by combining five different models including Naive Bayes, Multi-Layer Perceptron, Simple Logistic and Decision Stump with bagging and stacking techniques to classify the graduates’ employability and achieve 98.5243% precision which outperforms other selected classification models.

In summary, following the lead of the literature review that the predictive models selected by most researchers, classification models were to be primarily used for constructing the workplace analytics prediction in this research. Thus, the machine learning approach of this study include Decision Tree Analysis, Random Forest Model selection, Naive Bayes Classification Method, Support Vector Machines and lastly K-Nearest Neighbor Algorithms which were chosen to be developed for workplace preference analytics to find graduates’ ideal company type and sector.

III. METHOD

In this study, workplace preference analytics studied the graduate-preferred work environments via detailed data analysis and machine learning approach. Various variables such as the location of the company, sector of the job, and public or private entities are among the important factors in understanding their preferences and priorities in selecting their ideal workplace. By analyzing the information on graduates, a more comprehensive understanding of what attracts graduates to certain companies and what they value in a workplace would be made, which can inform the selection strategies for their career path. Therefore, a well-defined method and plan should be thoroughly outlined to

accomplish the study. a project framework with the flow and techniques of constructing the workplace preference analytics are proposed and explained in depth for each stage in this section, as shown in Figure 1.

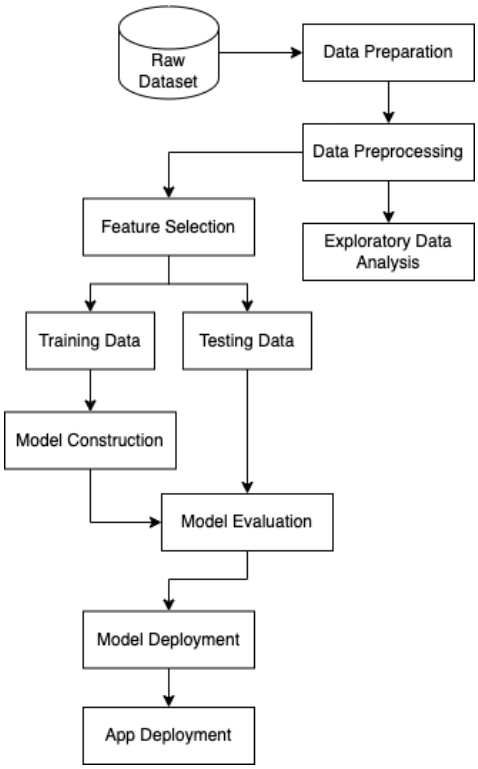

Figure 1: Flowchart of the Project Framework

A. Data Preparation

As shown in Figure 1, data preparation is the first process of the data mining approach to start with. Data collection is completed at the first stage since the present graduate dataset needed is sourced by a university as real-world data. Consequently, understanding the structure and definition of the data is essential for better analysis and optimization in the later phase. In this section, data gathering, merging, structuring and organizing are all included for constructing a well-structured dataset. In addition, the analysis work of this project is supplemented by analyzing two additional datasets from other sources which are the analytical datasets of graduate's living place and their current company location to further extract the characteristics specifically on location.

Firstly, a graduate dataset is the primary dataset used in this research is the graduate dataset. The dataset consists of 2382 profiles of graduates and comprises 100 features that provide insights into the graduates' academic profile, academic achievement, current job information and university-related opinion reviews. For instance, this dataset includes the graduate's course name, graduation date, CGPA (Cumulative Grade Point Average) score, MUET (Malaysian University English Test) results, current employment status, present workplace location, living location and more representing their previous profile as a student. These graduates are expected to complete their respective programmes on dates ranging from 16 October 2011 and 1 August 2021 (see Table 3).

Table 3: Partial Graduate Dataset

	Index	City name	Permanent address	Address latitude	Address longitude	Postal code	State name	Permanent city name
1	UG1	Melaka Tengah	NO. 9, JLN HULU LANGAT JAYA 2/1, TAMAN HULU LA…	3.0685514	101.7824036	43100	Selangor	Ulu Langat

2	UG2	Ulu Langat	NO 3 JALAN 9/5, TAMAN BUKIT MEWAH PHASE 9/10, M…	2.9739971	101.8112171	43000	Selangor	Ulu Langat
3	UG3	Melaka Tengah	26, JALAN KL 3/1B, TAMAN KOTA LAKSAMANA, SEKS 3...	2.1985218	102.2375032	75200	Melaka	Melaka Tengah
…	…	…	…	…	…	…	…	…
2380	UG2380	Putrajaya	1, JALAN P18H 1/7, PRESINT 18	2.9138407	101.6964668	62150	Wilayah Persekutuan Putrajaya	Putrajaya
2381	UG2381	Kuantan	NO 3 LORONG IM 5/15, BANDAR INDERA MAHKOTA	3.8221945	103.2887377	25200	Pahang	Kuantan
2382	UG2382	Gombak	NO 3, JALAN ANGGERIK, TAMAN UDA JAYA,	3.1581121	101.7656028	68000	Selangor	Gombak

Secondly, a geographically-focused analytical dataset namely residential geo-location dataset is constructed upon the existing main dataset by performing analytics on the living location of the graduates. The location information obtained includes the feature properties surrounding the graduates' living place, relative wealth index and population of nearby properties which extend a wider view of analysis for better prediction. After merging the newly constructed analytical dataset of graduate location with the main dataset, the new combination dataset eventually contains 2382 rows and 165 columns (Table 4).

Table 4: Residential Geo-Location Dataset

	latitude	longitude	KFC	7 Eleven Malaysia Sdn Bhd	Pizza Hut	McDonalds	Dominos	99 Speedmart Sdn Bhd	Dommal Food Services Sdn Bhd
0	1.352780	104.225439	1	0	0	0	0	0	0
1	1.361640	103.445397	0	0	0	0	0	0	0
2	1.366993	110.407094	0	0	0	0	0	0	0
…	…	…	…	…	…	…	…	…	…
2217	6.431306	100.198519	1	1	1	1	1	0	0
2218	23.748335	90.368817	0	0	0	0	0	0	0
2219	23.749648	90.380346	0	0	0	0	0	0	0
2220	23.764083	90.358209	0	0	0	0	0	0	0
2221	23.885942	45.079162	0	0	0	0	0	0	0

Thirdly, a second geographically-focused analytical dataset namely the workplace geo-location dataset was built on top of the primary dataset with the same method as the residential geo-location dataset. This analytical dataset is focusing on the current workplace location of the employed graduates. Similarly, this analytic acquires information on neighbouring properties of the graduates' present workplace, the relative wealth index, and the population of nearby properties surrounding the company they are working for. Finally, a new workplace analytical dataset is formed after merging the main dataset with the geographic data and it contains 2382 rows with 165 columns (Table 5).

Table 5: Workplace Geo-Location Dataset

	latitude	longitude	KFC	7 Eleven Malaysia Sdn Bhd	Pizza Hut	McDonalds	Dominos	Starbucks	Nandos
0	1.301813	103.849195	0	0	0	0	0	0	0

1	1.303643	103.854318	0	0	0	0	0	0	0
2	1.306984	103.829198	0	0	0	0	0	0	0
…			…	…	…	…	…	…	…
2217	6.116574	102.238834	1	0	1	1	1	0	1
2218	6.117832	102.24089	1	0	1	1	1	0	1
2219	6.141315	100.354009	1	0	1	0	1	0	0
2220	6.154821	100.371681	0	0	1	0	0	0	0
2221	6.327908	99.840714	1	0	1	0	0	0	0

B. Data Preprocessing

Data preprocessing plays a vital role in the data analysis pipeline, serving as an essential step after the collection of an adequate quantity of data from a specific target audience (Figure 1). It is the procedure of manipulating raw data into a more refined and structured form that is ready for precise analysis and modeling. It is crucial in data mining as it helps mitigate issues such as manual input errors, missing data, and redundant information, thereby improving the performance of predictive models. This involves a series of data cleaning and data transformation processes aimed at enhancing data quality by ensuring its completeness, correctness, and consistency. By undertaking these steps, the subsequent procedures can be facilitated more effectively.

In the first stage of data preprocessing, data transformation is implemented which used where the original data to be modified or converted to a suitable format for further analysis. The first case within this dataset is the column named “GOT” that stands for graduate on time contains 3 values which are “Yes”, “No” and “KIV”, which stands for “Keep in View” in cases when the total length of the term that graduate spent is shorter than expected. To standardize the values into only 2 categories with only “Yes” and “No”, the technique by grouping “KIV” with “Yes” is executed. Hence, students who have graduated earlier than expected can be included into the “Yes” category which still align with the notion of graduating on time. The analysis that relies on this column can be simplified and focuses on the binary categorization of on-time graduation (Yes) versus not on-time graduation (No).

Besides, the second problem can be seen from the column labeled “Subject Grades” which contains a bunch of subject results for each graduate record, made up of the SPM (Sijil Pelajaran Malaysia in Malay known as the Malaysian Certificate of Education) result of different subjects and separated with commas. This structure created difficulties on data modeling and analysis on the later stage. With the solution of splitting the data into separate columns, the data only align with the requirements of various modeling techniques such as the parameter of modeling in classification model. For instance, the English and Malay subjects namely “Bahasa Inggeris(SPM-2013):D, Bahasa Melayu(SPM-2013):E” can be separated into 2 columns which are “Bahasa Inggeris(SPM-2013)” and “Bahasa Melayu(SPM-2013)” with assigned value “D” and “E” respectively. This allows for easier integration of the data into modeling algorithms and lastly improves the accuracy and effectiveness of the analysis. There are 340 columns derived from the "Subject Grades" column, however 312 of them have 2000 or more blanks, meaning that about 80% of the data is missing. Therefore, "Subject Grades" was eliminated since it introduced noise into the data.

Furthermore, data cleaning is a critical initial step in the data analytics process that significantly impacts the quality and reliability of the dataset. It ensures the completeness and accuracy of the data by addressing unreasonable values and detecting errors through the cleaning of dirty data. The data type of each column is checked through exploratory data analysis to ensure it is compatible and suitable for further analysis. The first data cleaning approach was focused on the address data of the graduate is cleaned up by identifying and eliminating unwanted symbols and special characters which is a form of noise within the dataset. Unnecessary symbols and characters may not convey any meaningful information and may be misinterpreted while being analyzed and processed. In this dataset, there is a strange symbol and character like “_x000D_” which is not possible to be part of a regular address. The possibility of this term occurring might be the data transferred or processed between different platforms or systems that interpret newline characters differently. Therefore, replacing these characters with an empty string is performed for the data cleaning process.

Moreover, missing data is essential to be checked and handled by removing the incompleteness and misleading insights (Figure 2). There are 2 columns having completely null values which are "Discount" and "Loan". Hence, these 2 columns are removed since columns named sponsor category, sponsor detail and scholarship are indicating the similar meaning on the types of financial support that institution supplied to the graduate. In addition, through the exploratory data analysis, there are 4 graduate records that were discovered having missing data on part of the important information such as date of birth, gender and even begin date of admit term. In this case, these 4 graduate records are dropped which consider their academic record is deemed invalid since info of date of birth is blank.

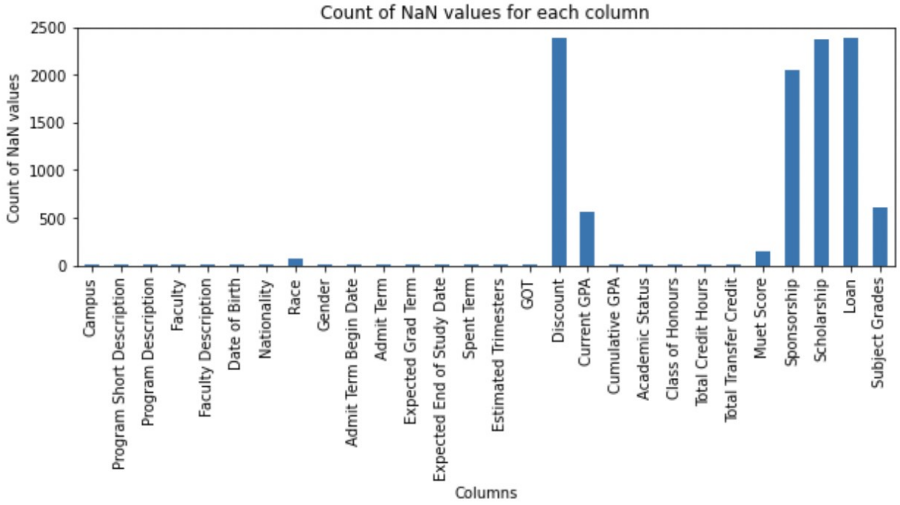

Figure 2: Null Values in Primary Dataset

Lastly, duplicated data is not detected in this dataset hence removal of redundant data is not needed. However, several duplicated features are found within this dataset, particularly in the form of columns that serve the same purpose for many variables. For instance, the results of Malaysian University English Test (MUET) are represented by a few columns named "muet", "Muet Score" and the "MUET(MUET)" that splitted out from originally "Subject Grades" in the data transformation phase previously. Instead of merging the values of 3 columns, only the value of the first "muet" column is preserved.

C. Exploratory Data Analysis

Exploratory data analysis (EDA) is an approach to analyze and summarize data for discovering patterns, relationships and insights with the use of various statistical and visualization methods. By examining the data structure through EDA, the analysis is conducted by understanding the main characteristics of the dataset, finding outliers and spotting possible trends or patterns. A few common techniques used in EDA include summary statistics, histograms, box plots, scatter plots and correlation analysis.

D. Feature Selection

Feature selection is used to determine which aspect of the given dataset is the most important. It is a process to select a subset of important variables or attributes from a larger set of available features within a dataset. Thus, the most informative features that contribute the greatest impact on the predictive performance of machine learning classifiers can be identified. It lowers the complexity of the predictive model by removing the least important features until the necessary number of significant features remains.

The 2 selected feature selection techniques used to rank the optimal feature are Boruta and Recursive Feature Elimination (RFE). The top 10 features are the highest scores on the ranking based on the evaluation of the importance of each feature. Table 6 and 7 shown below indicates the top-10 and bottom-10 features:

Table 6: Features Ranking using Boruta

No.	Top-10 Features	Score	Bottom-10 Features	Score
1	Jobstreet Company Industry	1.00	Work in Same Field of Learning	0.33
2	Award Degree	1.00	Nationality	0.30
3	Current Job Level	1.00	Part-Time Job	0.26
4	Cumulative GPA	1.00	ielts	0.26
5	Race	0.96	Entry Eligibility	0.19
6	Faculty	0.93	Study Method	0.15
7	Sub-area of Study Field	0.89	Scholarship	0.11
8	Sponsorship	0.85	oku	0.07
9	Main Job Scope	0.81	toefl	0.04
10	Specialized Study Field	0.78	Reason for being unemployed	0.00

Table 7: Features Ranking using RFE

No.	Top-10 Features	Score	Bottom-10 Features	Score
1	Jobstreet company industry	1.00	Work in Same Field of Learning	0.39
2	Sub-area of Study Field	1.00	Part-Time Job	0.35
3	Race	1.00	Entry Eligibility	0.30
4	Cumulative GPA	1.00	Study Method	0.26
5	Current Job Level	1.00	Nationality	0.22
6	Sponsorship	1.00	ielts	0.17
7	Award Degree	1.00	Scholarship	0.13
8	Faculty	1.00	Reason for being unemployed	0.09
9	Graduate On Time	0.96	toefl	0.04
10	Main Job Scope	0.91	oku	0.00

Tables 6 and 7 indicate the ranking from the most important features to the least impacting features selected by Boruta and RFE. The outcome of the top-10 selected features of both techniques are almost the same attributes but ranked differently.

E. Model Construction

For constructing workplace preference analytics, several classification algorithms are developed for the predictive models. The selected classification algorithms are Decision Tree, Random Forest, K-Nearest Neighbor, Naive Bayes and Support Vector Machine are used to identify desired company types and sectors based on their preferences and preferred location.

Decision Tree is a supervised learning algorithm that creates a flowchart-like model, where each internal node represents a feature, each branch represents a decision based on that feature, and each leaf node represents a predicted outcome or class label. Decision Trees are versatile, interpretable, and capable of handling both classification and regression tasks.

Random Forest is an ensemble learning method that combines multiple Decision Trees. It works by training several Decision Trees on different subsets of the training data and making predictions based on the average or majority vote of the individual trees. Random Forests can improve prediction accuracy, reduce overfitting, and handle high-dimensional data.

K-Nearest Neighbor (KNN) is a non-parametric algorithm used for both classification and regression tasks. It assigns a new data point to the class or predicts its value based on the majority vote or average of the K-nearest neighbours in the training data. KNN does not require training, but it can be computationally expensive for large datasets.

Naive Bayes is a probabilistic classifier based on Bayes' theorem with the assumption of independence between features. It calculates the posterior probability of a class given the feature values and predicts the class with the highest probability. Despite its simplicity, Naive Bayes performs well in many real-world applications and is particularly effective with high-dimensional data.

Support Vector Machine (SVM) is a powerful supervised learning algorithm used for both classification and regression tasks. It finds an optimal hyperplane that maximally separates the data points of different classes while maximizing the margin. SVM can handle linear and non-linear classification tasks by using different kernel functions. It is effective in high-dimensional spaces and can handle datasets with few instances.

F. Model Evaluation

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

Accuracy is one metric to evaluate a classification model's performance. Accuracy can be thought of as the percentage of correct predictions for the test data. It is a simple calculation by dividing the number of correct predictions by the number of total predictions as Equation (1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision as known as positive predictive value is the next place to be inspected in metric. Precision is the proportion of correctly predicted positive classifications from the cases that are predicted to be positive. In other words, precision measures how often a model predicts correctly in the positive class as Equation (2).

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall is also known as sensitivity or true positive rate. Recall is the proportion of correctly predicted positive classifications from the cases that are actual positives. In other words, recall measures how frequently a model predicts a positive when it is truly positive as Equation (3).

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1 score is a weighted average of precision and recall. Keeping an eye on the F1 score could be the optimal way to strike the best possible balance between recall and precision. In most cases, the F1 score is more beneficial than accuracy, particularly when dealing with uneven class distribution. Accuracy is useful when there are similar costs in FP and FN, but it's suggested to pay close attention to both precision and recall the difference between the cost of FP and FN are vary greatly. The formula used for F1 score is as Equation (4).

$$F1\ Score\ = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

IV. FINDINGS

The classification models were conducted to gain insights and draw meaningful conclusions from the data. The accuracy of the classification model indicates its effectiveness in accurately predicting the classes or labels of the data instances. This result demonstrates the model's ability to generalize well and make accurate predictions for new, unseen data points.

After a classification model has been developed, it must be evaluated to see how well it performs in making predictions on unseen data. To acquire a comprehensive image of the model's performance from multiple angles, it's always recommended to use a variety of approaches. Based on the study in section 2.5, multiple metrics and tools, including accuracy, precision, recall, and F1 score, can be used to assess the quality of a machine learning model. Therefore, a high-performing model can be chosen for the next process which is deployment. In this section, an accuracy comparison among different classifiers is conducted for three distinct prediction outcomes and analyzes the impact that the feature selection has on this framework.

A. Employment Status Prediction

The prior work in finding a suitable workplace for them is predicting the employment status of a graduate on the basis of their student profile. The outcome will identify whether the graduate is being employed which is a binary classification. The comparison of the accuracy, precision, recall and FI-score of the five classifiers was presented in the Table 8 to Table 10.

Table 8: Employment Status Prediction without Feature Selection

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Decision Tree	99.30	99.78	99.12	99.45
Random Forest	**99.40**	**100.00**	**99.32**	**99.66**
Support Vector Machine	70.70	69.65	100.00	82.11
K-Nearest Neighbor	98.50	100.00	97.79	98.89
Naive Bayes	99.30	99.78	99.12	99.45

Table 9: Employment Status Prediction using Boruta

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Decision Tree	99.30	99.77	99.12	99.44
Random Forest	99.40	99.76	99.30	99.53
Support Vector Machine	99.00	99.33	99.12	99.22
K-Nearest Neighbor	99.10	100.00	99.11	99.33
Naive Bayes	**99.40**	**100.00**	**99.30**	**99.65**

Table 10: Employment Status Prediction using RFE

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Decision Tree	**99.90**	**99.79**	**100.00**	**99.89**
Random Forest	99.70	99.78	99.78	99.78
Support Vector Machine	99.60	99.58	99.79	99.68
K-Nearest Neighbor	99.70	100.00	99.56	99.78
Naive Bayes	99.40	100.00	99.12	99.56

The finding of the comparison of Tables 8, 9 and 10 above is that the five predictive models are able to perform well without the need of using feature selection techniques, Boruta and RFE. As the result of the prediction without feature selection, the Random Forest model outperforms other classifiers with the highest F1-score which is 99.66% which is considered an excellent classifier. In addition, although Naïve Bayes with Boruta achieved the highest F1-score 99.65% among the other classifiers with Boruta which is 0.20% higher than the base model, there is no significant improvement in the metrics result of the other models with Boruta compared to the models without feature selection.

Besides, with the RFE feature selection approach, the F1-score only increased by 0.33%, hence the base model is sufficient to predict employment status in this research.

B. Company Type Prediction

For helping graduates find their desired workplace, the next stage is conducting the prediction based on their profiles and geographic analysis information. These models will address multiclass classification and make predictions about the type of company, such as whether they are suggested to go for local companies, state government, federal government and other forms of business. A variety of metrics results were compared as shown in the following tables (Table 11 to 13).

Table 11: Company Type Prediction without Feature Selection

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Decision Tree	62.50	43.69	62.47	50.94
Random Forest	**61.50**	**56.69**	**61.54**	**55.43**
Support Vector Machine	47.30	30.85	47.31	32.31
K-Nearest Neighbor	42.70	37.16	42.66	39.03
Naive Bayes	24.50	57.37	24.48	18.61

Table 12: Company Type Prediction using Boruta

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Decision Tree	62.50	43.69	62.47	50.92
Random Forest	**66.60**	**60.45**	**65.98**	**60.18**
Support Vector Machine	48.70	39.23	48.72	38.65
K-Nearest Neighbor	42.90	37.26	42.89	39.16
Naive Bayes	60.80	54.57	60.84	56.47

Table 13: Company Type Prediction using RFE

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Decision Tree	62.50	43.69	62.47	50.94
Random Forest	**65.70**	**60.25**	**65.73**	**59.89**
Support Vector Machine	48.30	38.44	48.25	38.34
K-Nearest Neighbor	42.00	35.92	41.96	38.04
Naive Bayes	61.80	55.82	61.77	57.52

In summary, the Random Forest model outperformed the other models in both experiments, regardless of whether feature selection was used. Although the greatest F1-score of 60.18% from Random Forest model with Boruta fell short of a good classifier which should have at least 75% of metrics, the significance of feature selection in enhancing the metrics was presented in this section. A rising F1-score of 5.38% is presented which indicates a good impact of feature selection in improving model performance. As a result, the Random Forest model with Boruta will be selected for company type prediction but additional data is required to train a superior classifier by addressing the issue of imbalance classes. The 8 classes that imbalance distributed on only 1400 graduates are causing the underfitting problem on the model which resulting a dropping of metric results.

C. Company Sector Prediction

After a company type has been suggested to the graduates, more options on the aspect of a workplace like company sector can be predicted to further assist them in making a career choice. Multiclass prediction in using their educational profile and analytical data on their location is conducted to predict the company sector that is suitable for them. For

instance, the company's area of expertise, whether it is IT, education, manufacturing, electronic or other company fields. The results of metrics among different classifiers are presented in the following table (Table 14 to 16) below:

Table 14: Company Sector Prediction without Feature Selection

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	26.10	13.07	26.11	16.24
Random Forest	**30.50**	**18.43**	**30.54**	**22.14**
Support Vector Machine	18.90	18.32	18.88	10.92
K-Nearest Neighbor	15.40	14.38	15.38	13.73
Naive Bayes	9.60	7.09	9.56	7.06

Table 15: Company Sector Prediction using Boruta

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	25.60	13.02	25.64	16.22
Random Forest	**30.80**	**19.93**	**20.77**	**22.91**
Support Vector Machine	21.00	20.28	20.98	17.61
K-Nearest Neighbor	17.20	16.45	17.25	15.29
Naive Bayes	14.70	18.87	14.69	10.51

Table 16: Company Sector Prediction using RFE

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	21.00	10.16	20.98	12.55
Random Forest	**22.10**	**10.95**	**22.14**	**14.04**
Support Vector Machine	17.00	14.39	17.01	14.82
K-Nearest Neighbor	9.10	9.90	9.09	8.25
Naive Bayes	1.40	3.02	1.40	1.00

As a result, the Random Forest algorithm again is trained as the most superior classifier among other classification models regardless of the existence of feature selection. Although the Random Forest model with Boruta achieved the highest F1-score of 22.91%, it is still considered the result of a very poor metric. This is due to the high imbalance class that consists of 27 classes of company sectors collected from only around 1400 graduates, hence more data is needed to be trained to address the underfitting problem. The feature selection does not affect the model performance significantly which can be observed in the small increased F1-score of only 0.77% on the company sector prediction using Boruta.

V. CONCLUSION

This research carries out workplace preference analytics with machine learning approaches to identify the employment status, preferred company type and sectors of graduates. Decision Tree, Random Forest, K-Nearest Neighbor, Naive Bayes, and Support Vector Machine are developed to build the predictive models for predicting an ideal workplace for graduates. For predicting the employment status of graduates, the base Random Forest model without feature selection is chosen as the top performer with the highest metrics results with accuracy of 99.40%, precision of 100%, recall of 99.32%, as well as F1-score of 99.66%. The base model is sufficient to provide a very accurate prediction even without the Boruta and RFE. Secondly, the best model for predicting the company type for the graduates is the Random Forest model using Boruta with an accuracy of 66.60%, precision of 60.45%, recall of 65.98% and F1-score of 60.18%. Thirdly, the most superior model of predicting the company sector for the graduates is again the Random Forest model using Boruta with an accuracy of 30.80%, precision of 19.93%, recall of 20.77% and F1-score of 22.91%. This indicates that the Random Forest algorithm was the most effective in predicting the employment status, and the appropriate company type and sector for graduates based on their profiles and geographic data. However, the need for the feature selection technique Boruta can be seen for the predictive models in predicting the company type and sector

but not for predicting employment status. For further improvement, the analytics system should be regularly updated with new data to ensure that the models are trained on the most recent information. Graduates' preferences and workplace trends may change over time, and the models should reflect these updates.

ACKNOWLEDGEMENT

A version of this paper was presented at the third International Conference on Computer, Information Technology and Intelligent Computing, CITIC 2023, held in Malaysia on 26th-28th July 2023.

REFERENCES

[1] R. Megasari, E. Piantari, R. Nugraha, "Graduates profile mapping based on job vacancy information clustering", 2020 6th international conference on science in information technology (icsitech), pp. 35-39), 2020.

[2] B. D. Puspasari, L. L. Damayanti, A. Pramono, A. K. Darmawan, "Implementation k-means clustering method in job recommendation system", 2021 7th international conference onelectrical, electronics and information engineering (iceeie), pp. 1-6, 2021.

[3] A. Olowolayemo, K. Harun, T. Mantoro, "University based job recommender alumni system", 2018 international conference on computing, engineering, and design (icced), pp. 212-217, 2018.

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

[5] Q. Zhou, F. Liao, L. Ge, J. Sun, "Personalized preference collaborative filtering: Job recommendation for graduates", 2019 ieee smartworld, ubiquitous intelligence computing, advanced trusted computing, scalable computing communications, cloud big data computing, internet of people and smart city innovation (smartworld/scalcom/uic/atc/cbdcom/iop/sci), pp. 1055-1062, 2019.

[6] O. Awujoola, P. O. Odion, M. E. Irhebhude, & H. Aminu, "Performance evaluation of machine learning predictive analytical model for determining the job applicants employment status", vol. 6, pp. 67-79, 2021.

[7] M. Nie, L. Yang, J. Sun, H. Su, H. Xia, D. Lian, K. Yan, "Advanced forecasting of career choices for college students based on campus big data", Frontiers of Computer Science, vol. 12, 2017.

[8] S. Li, Y. Chuancheng, W. Hongguo, D. Yanhui, "An employment recommendation algorithm based on historical information of college graduates", 2018 9th international conference on information technology in medicine and education (itme), pp. 708-711, 2018.

[9] H. Yu, Z.-q. Zhang, "The application of data mining technology in employment analysis of university graduates", 2018 ieee/acis 17th international conference on computer and information science (icis), pp. 846-849, 2018.

[10] D. Chakraborty, M. S. Hossain, M. S. Arefin, "Demand analysis of cse graduates of different universities in job markets", 2019 international conference on electrical, computer and communication engineering (ecce), pp. 1-6, 2019.

[11] Z. Chen, W. Liang, X. Gao, Z. Zhou, M. Wu, "Research on the accurate recommendation management system for employment of college graduates on Hadoop", 2019 5th international conference on big data and information analytics (bigdia), pp. 19-22, 2019.

[12] M. Jiang, Y. Fang, H. Xie, J. Chong, M. Meng, "User click prediction for personalized job recommendation", World wide web 22, pp. 325-345, 2019.

[13] M. Kahn, T. Gamedze, J. Oghenetega, "Mobility of sub-saharan africa doctoral graduates from south african universities—a tracer study", International Journal of Educational Development, vol. 68, pp. 9-14, 2019.

[14] K. Kusnawi, J. Ipmawati, D. Kusumandaru, "Decision support system employee recommendation using fuzzy sugeno method as a job search service", 2019 international conference on information and communications technology (icoiact), pp. 539-542, 2019.

[15] A. Nigam, A. Roy, H. Singh, H. Waila, "Job recommendation through progression of job selection", 2019 ieee 6th international conference on cloud computing and intelligence systems (ccis), pp. 212-216, 2019.

[16] L. G. Rodriguez, E. P. Chavez, "Feature selection for job matching application using profile matching model", 2019 ieee 4th international conference on computer and communication systems (icccs), pp. 263-266, 2019.

[17] Z. Yang, S. Cao, "Job information crawling, visualization and clustering of job search websites", 2019 ieee 4th advanced information technology, electronic and automation control conference (iaeac), vol. 1, pp. 637-641, 2019.

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

[19] H. Fan, "A prediction model of college students' employment based on data mining", 2020 13th international conference on intelligent computation technology and automation (icicta), pp. 549-552, 2020.

[20] D. Hooshyar, M. Pedaste, Y. Yang, "Mining educational data to predict students' performance through procrastination behavior", Entropy, vol. 22(1), 2020.

[21] R. S. Kumar, N. Prakash, S. Anbuchelian, "Prediction of job openings in it sector using long short -term memory model", 2020 fourth international conference on i-smac (iot in social, mobile, analytics and cloud) (i-smac), pp. 945-953, 2020.

[22] N. Gunarathne, S. Senaratne, R. Herath, "Addressing the expectation–performance gap of soft skills in management education: An integrated skill-development approach for accounting students", The International Journal of Management Education, vol. 19(3), p. 100564, doi: https://doi.org/10.1016/j.ijme.2021.100564, 2021.

[23] B. Heriyadi, "Tracer study analysis for the reconstruction of the mining vocational curriculum in the era of industrial revolution 4.0", Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, pp. 3013-3019, 2021.

[24] N. Premalatha, & S. Sujatha, "An effective ensemble model to predict employment status of graduates in higher educational institutions", 2021 fourth international conference on electrical, computer and communication technologies (icecct), pp. 1-4, 2021.

[25] M. Rahman, M. Asadujjaman, "Multi-criteria decision making for job selection", 2021 international conference on decision aid sciences and application (dasa), pp. 152-156, 2021.

[26] S. R. Nudin, B. Warsito, A. Wibowo, "Impact of soft skills competencies to predict graduates getting jobs using random forest algorithm", 2022 1st international conference on information system information technology (icisit), pp. 49-54, 2022.

[27] Y. Bharambe, N. Mored, M. Mulchandani, R. Shankarmani, S. G. Shinde, "Assessing employability of students using data mining techniques", 2017 international conference on advances in computing, communications and informatics (icacci), pp. 2110-2114, 2017.

[28] M. M. Almutairi, M. H. A. Hasanat, "Predicting the suitability of is students' skills for the recruitment in saudi arabian industry", 2018 21st saudi computer society national computer conference (ncc), pp. 1-6, 2018.

[29] M. Y. Arafath, M. Saifuzzaman, S. Ahmed, S. A. Hossain, "Predicting career using data mining", 2018 international conference on computing, power and communication technologies (gucon), pp. 889-894, 2018.

[30] J. Martinez-Gil, B. Freudenthaler, T. Natschläger, "Recommendation of job offers using random forests and support vector machines", 2018.

[31] A. Nachev, T. Teodosiev, "Analysis of employment data using support vector machines", International journal of applied engineering research issn, vol. 13, pp. 13525-13535. 2018.

[32] R. Amalia, A. Wibowo, "Prediction of the waiting time period for getting a job using the naive bayes algorithm", International research journal of advanced engineering and science, vol. 5, pp. 225-229, 2020.

[33] A. Daharwal, Prof. S. Gore., A. Bhagwat., S. Dethe., S. Chavan., "Career guidance system using machine learning for engineering students (cs/it)", International research journal of engineering and technology (irjet), vol. 7, pp. 3417-3420, 2020.

[34] K. Jamal, R. Kurniawan, I. Husti, Zailani, M. Z. A. Nazri, J. Arifin, "Predicting career decisions among graduates of tafseer and hadith", 2020 2nd international conference on computer and information sciences (iccis), pp. 1-4, 2020.

[35] T. V. Huynh, K. V. Nguyen, N. L.-T. Nguyen, A. G.-T. Nguyen, "Job prediction: From deep neural network models to applications", 2020 rivf international conference on computing and communication technologies (rivf), pp. 1-6, 2020.

[36] R. G. Angesti, A. Kurniawati, H. D. Anggana, "Prediction of the telkom university's undergraduates waiting period for getting a job using the cart algorithm", 2021 4th international conference of computer and informatics engineering (ic2ie), pp. 135-140, 2021.

[37] B. M. D. E. Bannaka, D. M. H. S. G. Dhanasekara, M. K. Sheena, A. Karunasena, N. Pemadasa, "Machine learning approach for predicting career suitability, career progression and attrition of it graduates", 2021 21st international conference on advances in ict for emerging regions (icter), pp. 42-48, 2021.

[38] M. Sharma, S. Joshi, S. Sharma, A. Singh, R. Gupta, "Data mining classification techniques to assign individual personality type and predict job profile", 2021 9th international conference on reliability, infocom technologies and optimization (trends and future directions) (icrito), pp. 1-5, 2021.

[39] J. Zhu, G. Viaud, C. Hudelot, "Improving next-application prediction with deep personalized-attention neural network", 2021 20th ieee international conference on machine learning and applications (icmla), pp. 1615-1622, 2021.

[40] S. Vassef, R. Toosi, M. A. Akhaee, "Job title prediction from tweets using word embedding and deep neural networks", 2022 30th international conference on electrical engineering (icee), pp. 577-581, 2022.