Journal of Telecommunication, Electronic and Computer Engineering
ISSN: 2180 – 1843 e-ISSN: 2289-8131 Vol. 16 No. 3 (2024) 1-7
jtec.utem.edu.my
DOI: https://doi.org/10.54554/jtec.2024.16.03.001

Malaysian Community College Graduates Employability Prediction Model Using Machine Learning Approach

Azida Mansor[1] and Zuraini Othman[2*]
[1]*Department of Information and Communication Technology, Politeknik Ungku Omar, 31400 Ipoh, Perak, Malaysia.*
[2]*Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia.*

Article Info

Article history:
Received Mar 29th, 2024
Revised Jul 23rd, 2024
Accepted Sep 2nd, 2024
Published Sep 30th, 2024

Index Terms:
Employability model
Machine Learning
Logistic Regression
Neural Networks
Random Forest

Abstract

Community College, as TVET institution under the Ministry of Higher Education, offers industry-relevant skills training to ensure graduates' employability in the global labor market. However, producing graduates who meet industry demands remains a challenge, and industries continue to face difficulties in obtaining skilled graduates. In Malaysia, there is limited research on predictive models for employability rates among graduates of TVET Malaysian institutions. This research utilizes Python to investigate the significant factors influencing employability among Malaysian Community College graduates, determined by a specific indicator of whether they will be employed. Our contribution lies in developing an accurate employability prediction model using machine learning algorithms such as Logistic Regression, Neural Networks, and Random Forest. The dataset used consisted of 10,427 instances and 14 attributes, from which six significant factors were identified. Among the models, Random Forest outperformed the other machine learning models, and hyperparameter tuning using RandomizedSearch further improved the accuracy of the model to 84.8%. This study aims to identify the most accurate and interpretable model, providing valuable insights for educational institutions to enhance their employability strategies.

*Corresponding Author: zuraini@utem.edu.my

I. INTRODUCTION

The Ministry of Higher Education aims to produce well-rounded and resilient graduates equipped with the knowledge and skills needed to thrive in a rapidly changing world [1]. Technical and Vocational Education (TVET) plays a crucial role in providing graduates with employability skills that are demanded across various industries. The Malaysian Government has invested substantial funding in public institutions to train TVET graduates to meet industry requirements and enhance their employability. Despite these efforts, the youth unemployment rate in Malaysia continues to rise, creating a highly competitive job market for graduates [2]. This study explores the challenges related to graduates' employability in Malaysia [3].

One of the main issues in the developing countries is that higher education programs often do not adequately prepare graduates for the job market. Research by [4] highlighted the importance of TVET in equipping students with practical skills needed for manual tasks. However, a gap still exists between the skills students acquire and the needs of the industry, particularly in Malaysia. The TVET curriculum is often outdated and lacks industry perspectives, which leads to a decline in student competencies and contributes to high unemployment rates. While previous studies have identified various factors that influence graduates' employability, research specifically focused on TVET programs is limited. For example, [5] developed a predictive model to assess TVET graduates' employability using machine learning classification algorithms. Additionally, research by [6] demonstrated that the Decision Tree algorithm yielded high accuracy rate and was straightforward to construct and interpret. These findings are now being used to optimize the model's hyperparameters and further improve its performance accuracy.

Data mining techniques were widely used in the field of Educational Data Mining, especially for classification tasks [7] [8]. This study examines the usefulness of various supervised machine learning algorithms in predicting graduate employability [9]. The findings suggest that communication skills, ethical behavior, and socio-economic, academic, and institutional factors significantly influence graduates' employability [10]. Among the algorithms used in these studies are Logistic Regression, Decision Trees, Random Forests and the unsupervised clustering (K-Means) algorithm, with each algorithm showing varying degrees of predictive accuracy [11]. Additional variables for further research have been proposed, such as student demographics, academic characteristics, and student satisfaction with university facilities (e.g., library and counseling service) in relation to graduate student employability predictions [12].

Machine learning offers the advantage of constructing models using both categorical and numerical predictors by analyzing linear and non-linear correlations between variables, as well as the relevance of each predictor [13]. In this study, three machine learning algorithms: Logistic Regression, Random Forest, and Neural Networks are utilized, and hyperparameter tuning is applied to improve the performance of the Malaysian Community College Graduates Employability Model. Logistic Regression is a technique that evaluates the impact of independent variables on a binary dependent variable [14]. Random Forest is an ensemble learning technique that builds multiple unpruned classification trees using bootstrap sampling, resulting in highly effective classification models [15]. Neural Networks, on the other hand, analyze complex, non-linear data using computational pattern recognition techniques [11]. Hyperparameter tuning is performed using grid search and random search to optimize the model's performance by finding the best hyperparameter settings, which control the model's behavior during training. Random search is found to be more efficient than grid search, particularly for tuning Neural Networks. The study will focus on Random Forest hyperparameter tuning, as it is expected to provide the highest accuracy [16].

This study aims to identify strategies to improve the employability of Malaysian Community College graduates by identifying significant factors that enhance their competency. The study will compare several machine learning classification algorithms, including Logistic Regression, Neural Networks, and Random Forest, to develop an accurate model for predicting graduate employability. This study aims to assist educational institutions in producing industry-ready graduates by providing insights and recommendations, Community College, as part of the Technical and Vocational Education and Training (TVET) institutions, will play a key role in encouraging students to develop the technical skills, knowledge, and social competencies needed to contribute effectively to various industries.

II. MATERIALS AND METHODS

This study aims to identify the key factors and methods employed by Malaysian Community College to enhance their graduates' competency levels and propose strategies to improve employability, ensuring that graduates secure jobs that match their qualifications. By comparing various machine learning classification methods, including Logistic Regression, Neural Networks, and Random Forest, this study intends to develop an accurate graduate employability prediction model. This model will predict whether a graduate is likely to be employed or unemployed. A wide range of evaluation metrics was used to validate the effectiveness of the classification algorithms. Additionally, hyperparameter tuning was applied to optimize the model's performance, as setting the correct combination of hyperparameters is crucial for maximizing the potential of the model.

The CRISP-DM (Cross Industry Standard Process for Data Mining) methodology is used to guide the research design for this study. The programming language Phyton is used to classify the data and evaluate empirical models during the experimental phase. The hardware setup includes an Intel Core i7-9700K processor and an NVIDIA GeForce RTX 2070 GPU, which is particularly used for training the Neural Network.

All the development and testing of the classification model will follow the Data Mining process to ensure high prediction accuracy. The study divides the development cycle into five major phases to perform the experimental research design: data collection, data preparation, data modelling, data evaluation and hyperparameter tuning (see Figure 1).

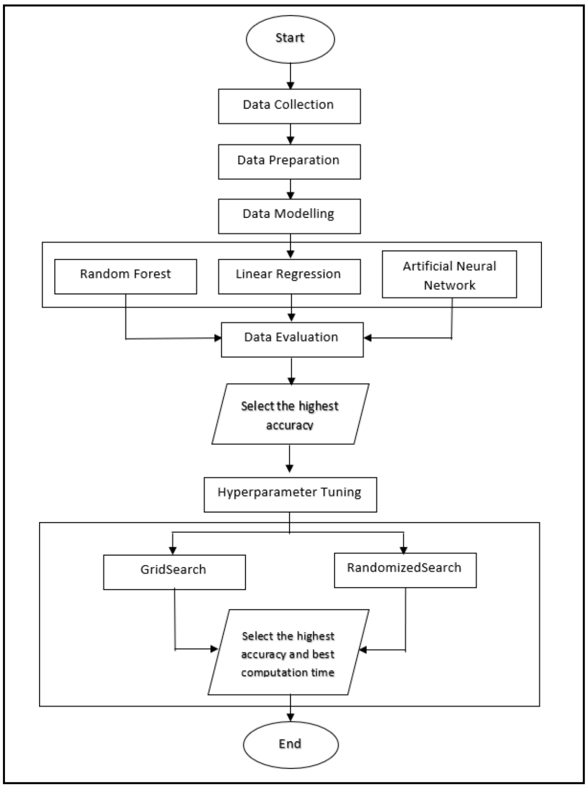

Figure 1. Research design flowchart

A. Data Collection

The data for this study was obtained from the TVET Graduate Tracer Study System (SKPG-TVET), a survey conducted by the Ministry of Higher Education among Community College graduates. The primary objective of this survey is to gather information on the employment status of graduates and obtain feedback on various aspects of their educational institutions. This dataset is particularly valuable as it provides a correlation between graduates' qualifications, skills and their employment status. The findings from this survey will play critical role in shaping future planning and enhancing different areas of the local TVET higher education administrative structure. Additionally, the survey serves as an essential tool for evaluating whether Malaysian higher education is adequately meeting workforce demands across various sectors, including technical, management, and social sciences. The dataset used in this study consists of data collected in 2019, comprising 11, 000 instances and 174 attributes compiled into a single worksheet.

B. Data Preparation

The process of preparing data for machine learning involves several steps, including parsing, cleaning, transformation, and pre-processing. Raw data often contains missing values, outliers, and irrelevant information, making data cleaning essential for removing these issues. Exploratory Data Analysis (EDA) is used to identify relationships between variables and detect class imbalances. After the data is cleaned, feature engineering and selection are performed to refine the features that are most important for the model. Label encoding is used to convert categorical features into numerical values, and feature selection is done using metrics such as chi-squared to determine which subset of features is most effective for predicting the dependent variable. In this study, the most relevant features were selected from the Tracer Study dataset to improve the accuracy of the classification algorithms.

C. Data Modelling

The research employs a supervised learning approach, specifically a classification model, to predict outcomes using the structured Malaysia Community College Graduates Employability dataset. The model development consists of two stages: the Training and Testing Phase, followed by the Prediction using Classification Algorithms.

In the Training and Testing Phase, the dataset is split into two parts: 80% is allocated for training, and 20% is reserved for testing. In the Prediction Phase, three machine learning classification algorithms, namely Logistic Regression, Random Forest, and Artificial Neural Network, are used to predict the employability of graduates. These algorithms were chosen based on recommendations from previous studies, as discussed earlier, and their appropriateness depends on the characteristics of the dataset.

After selecting the appropriate algorithm, the model is trained using the training dataset. Once the training is completed, the model is ready for testing using the testing dataset. The training dataset is used to fit and build the machine learning model, while the testing dataset is used to evaluate the fit and performance of the machine learning model for each algorithm applied.

Random Forest is an ensemble learning-based algorithm that multiple decision trees on different subsets of the dataset and averages their predictions to enhance the dataset's prediction accuracy. Logistic Regression is a supervised learning technique that uses independent factors to predict a categorical dependent variable, where the outcome is a binary (can only be between '0' and '1'). Artificial Neural Network is a processing unit consisting of neurons that tries to replicate the structure and behavior of the natural neuron.

In using these algorithms to solve classification problems, the conventional machine learning pipeline is followed. This pipeline involves importing the required classes, specifying the necessary parameters, creating and training the model using the chosen algorithm on a training dataset, predicting the model using the testing dataset, and finally evaluating the model’s performance.

D. Data Evaluation

In this study, three classification algorithms are applied to evaluate the performance of the graduates' employability model. The study focuses on binary classification, where graduates are classified as either employed or unemployed. The model’s performance is evaluated using accuracy and the confusion matrix. The confusion matrix is a contingency table that represents the relationship between the predicted and actual class classifications. The values along the diagonal represent correct classifications, while the off-diagonal values represent misclassifications.

Several evaluation metrics, derived from the confusion matrix, were used to assess the model’s performance, particularly metrics that are sensitive to changes in classification outcomes and take into account the financial cost of incorrect prediction. These metrics include recall, precision, specificity, false positive rate, accuracy, and the area under the ROC curve (AUC). Recall measures the model's ability to correctly identify employed graduates. Precision evaluates how many of the predicted employed graduates were accurately classified. Specificity indicates the model’s ability to correctly identify unemployed graduates, essentially rejecting incorrect classifications. The false positive rate measures the proportion of incorrect positive predictions, where a lower rate is preferable. The area under the ROC curve assesses the model’s overall ability to distinguish between employed and unemployed graduates.

E. Proposed Work for Hyperparameter Tuning

This section discusses hyperparameter tuning in machine learning, with a particular focus on the Random Forest algorithm. The aim of hyperparameter tuning is to identify the optimal set of hyperparameters that maximize the model’s accuracy when evaluated on a validation set. Two techniques commonly used for hyperparameter tuning are Grid Search and Random Search.

In Grid Search, a sparse parameter grid is defined, and the performance of each model configuration is evaluated using a grid search algorithm. The model that performs best is then selected for predicting the test dataset. The hyperparameters tuned using Grid Search for the Random Forest algorithm include n_estimators, max_depth, max_features, min_samples_leaf, min_samples_split, n_jobs, and Bootstrap.

In Random Search, random combinations of hyperparameters are selected, and each combination is used to train and score the model. Similar to Grid Search, the best-performing model configuration is chosen to predict the test dataset. The hyperparameters tuned in Random Search for Random Forest include the same parameters used in Grid Search.

Both Grid Search and Random Search are effective methods for optimizing the hyperparameters of a machine learning algorithm. The choice between the two approaches typically depends on the computational resources available and the number of hyperparameters to be tuned.

III. RESULTS AND DISCUSSION

This research aims to identify the factors and strategies used by Malaysian Community Colleges to enhance graduates' competency levels and improve their employability. The study also focuses on constructing a predictive model using machine learning classification techniques to determine whether a graduate is likely to be employed. Various evaluation metrics are used to validate the performance of the classification algorithms, and hyperparameter tuning is used to optimize the model's accuracy.

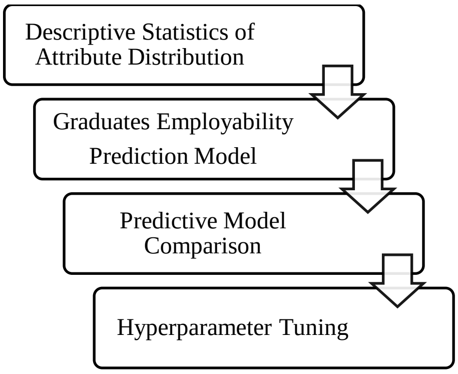
Descriptive Statistics of Attribute Distribution
Graduates Employability Prediction Model
Predictive Model Comparison
Hyperparameter Tuning

Figure 2. Result and analysis flowchart

Figure 2 highlights the four key findings of the research. The figure highlights the importance of aligning the project with its objectives and presents a flowchart illustrating the result analysis process of this research. The first step involves providing distribution frequencies for 13 dependent variables and one independent variable. The machine learning methodology used for the prediction process is divided into three primary steps. Firstly, the dataset is divided into training and testing datasets. Second, feature selection is performed to minimize the number of input variables. The third step involves performance measure, namely the accuracy and confusion matrix, used to compare techniques and identify the best-performing model. The performance is further improved by incorporating additional parameters into the existing approach.

A. Descriptive Statistics of Attribute Distribution

This section provides the distribution frequencies for 14 attributes: status, e_jantina, e_umur, e_keturunan, e_kampus, e_program, CGPA, kurikulum, latihan_industri, subjek_wajib_institusi, Maklumat_kerjaya, bantuan_pekerjaan, tenaga_pengajar and prasarana. The term "distribution frequencies" refers to a preliminary analysis aimed at better understanding the characteristics of the cleaned graduate employability dataset through various visualization methods. Using Exploratory Data Analysis (EDA), it was essential to identify correlations and trends that could influence the overall analysis. After this step, feature engineering is applied, followed by the initiation of machine learning modeling. This process is crucial for developing and refining the feature selection method used in the graduate employability prediction model. In this research, PowerBI is used to extract, convert, and load the data, facilitating the EDA procedures.

B. Graduates Employability Prediction Model

In this study, the classification model, a supervised learning approach, is used to predict outcomes based on the structured Malaysian Community College Graduates Employability dataset. This section explains the process of developing the graduates' employability prediction model, focusing on three key stages: Training and Testing, Feature Selection, and Prediction Using Classification Algorithms.

1) Training and Testing Phase

The dataset is split into two groups: the training dataset, which makes up 80% proportion, and the testing dataset, which accounts for the remaining 20% of the entire dataset. This results in 8,341 samples for training and 2,086 samples for testing.

The training dataset shows that the probabilities of graduates being employed or unemployed (dependent variables) are not equal, indicating an imbalance dataset. Specifically, the majority class (*Bekerja* = 6,235) is much larger than the minority class (*Tidak Bekerja* = 2,106). In this situation, the predictive model developed using conventional machine learning algorithms could be biased, leading to inaccurate prediction and unsatisfactory classifiers when dealing with imbalanced datasets. Figure 3 shows the graphical representation of the imbalance target, while Figure 4 shows the target class after applying the resampling technique to address this issue.

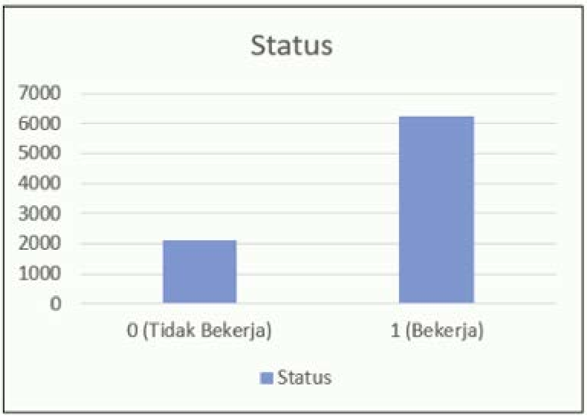

Figure 3. Imbalance target class

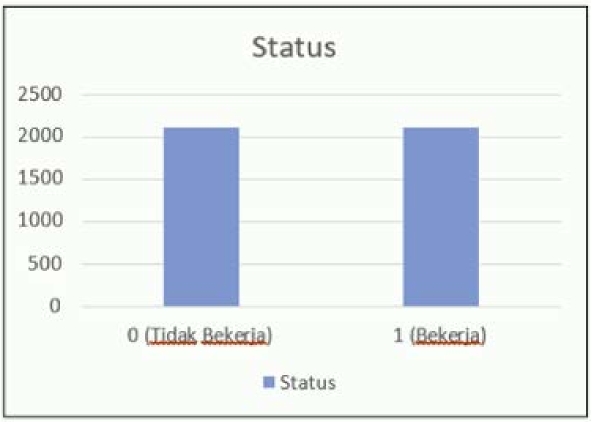

Figure 4. Target class after resample

2) Feature Selection

Next, feature selection is performed after the dataset split and combined with random under-sampling. Feature selection is the process of reducing the number of input variables when developing a predictive model. This step is crucial as it helps reduce the computational cost of modelling and, in some cases, improves the model's performance. Statistical-based feature selection methods evaluate the relationship between each input variable and the target variable using statistical techniques, selecting only those input variables that show the strongest correlation with the target variable. In this study, chi- square feature selection is used, as it is the most common correlation measure for categorical data.

A bar chart depicting the feature importance scores for each input variable is created, and Table 1 shows the p-values of the variables. P-value less than 0.05 (5% level significance) indicate that six independent variables are significant, as listed below.

Table 1
Chi-Square feature selection results

Most Significant	Variables	P-Values
1	***Kurikulum***	**1.350760e^{-120}**
2	***Subjek_wajib_institusi***	**5.324058e^{-41}**
3	***E_program***	**1.907199e^{-31}**
4	***E_jantina***	**8.914037e^{-08}**
5	***E_umur***	**1.670390e^{-06}**
6	***E_kampus***	**6.411564e^{-04}**
7	*Prasarana*	3.309050e^{-01}
8	*Latihan_industri*	6.144323e^{-01}
9	*Tenaga_pengajar*	6.185927e^{-01}
10	*E_keturunan*	7.265053e^{-01}
11	CGPA	8.470563e^{-01}
12	*Bantuan_pekerjaan*	8.777639e^{-01}
13	*Maklumat_kerjaya*	8.864422e^{-01}

Based on the Chi-Square feature selection results, this research tested the significance of the 13 independent variables. The result show that the non-significant variables are *'maklumat_kerjaya', 'bantuan_pekerjaan', 'CGPA', e_keturunan', 'tenaga_pengajar', 'latihan_industri'* and *'prasarana'.*

In this section, the best attributes were selected from the Tracer Study dataset to determine which subset of attributes most effectively predicts the accuracy of the classification algorithm. This research identified six significant variables, which are *e_kampus, e_umur, e_jantina, e_program, subjek_wajib_institusi* and *kurikulum*. These six significant factors were chosen for inclusion in the graduates' employability prediction model to address the first research question.

3) *Prediction Model Comparison using Classification Algorithms*

The experiment was designed to compare the performance of the classifiers, with the goal of selecting the best predictive model for the graduates' employability prediction. Figure 5 depicts the output of the ROC curve, showing the results of three prediction models: Random Forest, Logistic Regression, and Artificial Neural Network. In terms of accuracy, Random Forest achieved the highest score at 80.7%, followed by Logistic Regression at 77.6%, and Artificial Neural Network at 72.8%.

The Random Forest also provided the largest area under the ROC curve (AUC) with a value of 79.4%, while Logistic Regression followed with 75.8%, and Artificial Neural Network produced the lowest AUC at 72.8%. The AUC ranges between 0 and 1, where 0 indicates that the model performs no better than a random classifier and 1 indicates that the model is a perfect classifier. Based on these results, we can infer that the Random Forest model, with its AUC score closest to 1, outperforms the other algorithms that yield lower AUC values.

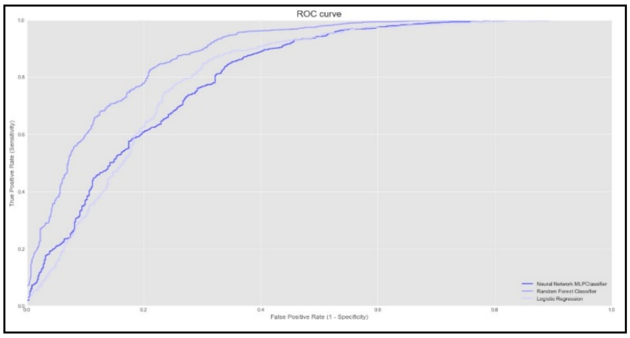

Figure 5. ROC Curve for all three algorithms

As shown in Table 2, the Random Forest classifier is the best at predicting graduates' employability, with an accuracy of 80.7% and an AUC of 79.4%, both of which are higher than the other algorithms. Random Forest outperformed the other machine learning models, addressing the second research question: which classification algorithms are suitable for developing a model of graduates' employability. Compared to Logistic Regression and Decision Trees, the results show that the Random Forest Classifier provides the best predictions of graduate students' employability, with an accuracy of 98% [11].

Table 2
Graduates employability prediction model comparison

Measures	Random Forest	Logistic Regression	Artificial Neural Network
Accuracy	0.807	0.776	0.728
Recall	0.823	0.797	0.727
Precise	0.905	0.885	0.879
AUC	0.794	0.758	0.728

4) *Proposed Hyperparameter Tuning*

This paper applies GridSearch and RandomizedSearch techniques for hyperparameters tuning on top-performing machine learning algorithms. Random Forest, which achieved the highest accuracy for classification problems, was further optimized using these techniques. The GridSearch and RandomizedSearch functions were then used to find the best combination of hyperparameters, followed by two additional runs. Figure 6 shows the confusion matrices for the three distinct hyperparameters configurations used in the Random Forest algorithm, while Table 3 displays the optimized hyperparameters. Additionally, Table 4 compares the computation times of GridSearch and RandomizedSearch..

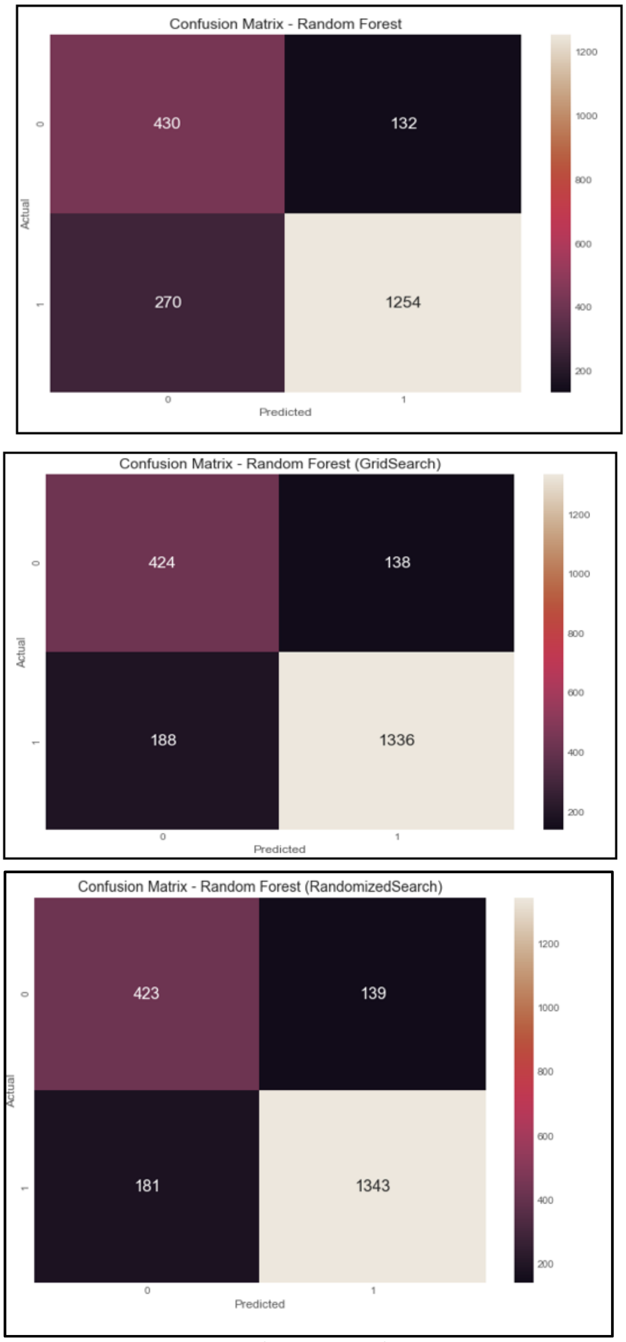

Figure 6. Confusion matrix for Random Forest

Based on the hyperparameter combinations, the number of true positives increased from 1,254 to 1,336 using GridSearch, and further to 1,343 using RandomizedSearch. This demonstrates that hyperparameters tuning improves the model's ability to correctly predict true positives, meaning the algorithm becomes more accurate in predicting employed graduates when they are actually employed.

Although the second combination better predicted the number of false negatives, reducing them from 270 to 188 (GridSearch) and 181 (RandomizedSearch), some predictions errors still occurred. In these cases, the model predicted a negative value (*Tidak Bekerja*), while the actual value was positive (*Bekerja*). This shows that hyperparameters tuning helps reduce miscalculation and improve the model's overall prediction.

The results in Table 3 demonstrates that optimized hyperparameter combinations significantly improve the model's performance compared to the default configurations. GridSearch increased accuracy by almost 4.0 percentage points, while RandomizedSearch improved it by 4.1 points.

Table 3
Optimal hyperparameters and accuracies for Random Forest algorithms

	Hyperparamaters	Accuracy
Default	{n_estimators= 100, max_depth=None}	80.7%
GridSearch	{'bootstrap': True, 'max_depth': 16, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 256}	84.7%
RandomizedSearch	{'n_estimators': 146, 'min_samples_split': 2, 'min_samples_leaf': 2, 'max_features': 'auto', 'max_depth': 16, 'bootstrap': True}	84.8%

Table 4
Comparison of computation time of GridSearch and Randomized

Classification Algorithm	With GridSearch	With Randomized
Random Forest	1.0s	0.5s

With GridSearch, the researcher set up a grid of these hyperparameter values and trained a model for each combination. This method tests all possible data combinations, which is not an ideal or efficient method. Moreover, this method is expensive in terms of computing power and time-consuming.

On the other hand, RandomizedSearch creates a grid of hyperparameter values and selects random combinations to train the model. This helps to control the number of parameters tested. Furthermore, the number of iterations of the search algorithm is based on available time and resources, making RandomizedSearch a more cost-effective and efficient option than the GridSearch.

As shown in Table 4, RandomizedSearch took only 0.5 seconds compared to 1.0 second for GridSearch, demonstrating its superior efficiency. Therefore, in response to the third research question, the effect of Random Forest classification algorithms with Hyperparameter optimization on the employability model of graduates has been evaluated. The performance measures discussed above show that RandomizedSearch outperform GridSearch in terms of both accuracy and computation time.

IV. CONCLUSIONS

As a conclusion, by comparing machine learning classification methodologies such as Logistic Regression, Neural Networks, and Random Forest, this study developed a machine learning model to create an accurate graduate employability prediction model. The findings indicate that two elements are crucial in determining the satisfaction of Community College graduates: the curriculum and the mandatory courses provided by the community colleges. These elements greatly contribute to the students' preparedness before entering the actual world of work.

The model was designed to predict whether a graduate would be employed or unemployed. The performance of each model was evaluated to determine the best model, with Random Forest achieving the highest accuracy for employability prediction. The GridSearch and RandomizedSearch strategies were used to tune the hyperparameters of the top machine learning algorithms. After optimizing the model parameters, Random Forest's accuracy increased from 80.7% to 84.7% using GridSearch and further to 84.8% using RandomizedSearch. This demonstrates that hyperparameter adjustment, particularly with Random Forest, as recommended by prior research, provides the highest level of accuracy. This research provides a robust framework for educational policymakers and institutions to enhance their strategies, ultimately improving graduate employability rates and aligning educational outcomes with labor market demands. Future work could build on these findings by exploring additional features and incorporating more advanced machine learning techniques to further refine the predictive models.

CONFLICT OF INTEREST

Authors declare that there is no conflict of interests regarding the publication of the paper.

AUTHOR CONTRIBUTION

The authors confirm contribution to the paper as follows: study conception and design: Azida Mansor and Zuraini Othman; data collection: Azida Mansor; analysis and interpretation of findings: Azida Mansor and Zuraini Othman; draft manuscript preparation: Azida Mansor. All authors had reviewed the findings and approved the final manuscript.

REFERENCES

[1] Ministry of Education Malaysia (MoE), "Malaysia education blueprint 2015-2025 (higher education)," Minist. Educ. Malaysia, vol. 2025, pp. 40, 2015.

[2] Department of Statistics Malaysia, "Department of Statistics Malaysia Official Portal," Accessed: Jul. 17, 2021, [Online] Available: https://www.dosm.gov.my/v1/index.php?r=column/cthemeByCat&cat=124&bul_id=a09WTHNBQVpVcHFiZkNTaEZnTHF3UT09&menu_id=Tm8zcnRjdVRNWWlpWjRlbmtlaDk1UT09.

[3] A. P. Tiwari, "Students' Enrollment decision in technical education in community schools modality," Social Inquiry: Journal of Social Science Research, vol. 5, no. 1, pp. 91-105, 2023, doi:10.3126/sijssr.v5i1.65413.

[4] Z. Buang, M. M. Mohamad, A. Ahmad, and N. Yuniarti, "The earnings and employment of community colleges' graduates: occupational field and gender analysis," J. Pendidik. Teknol. dan Kejuru., vol. 26, no. 1, pp. 11–17, 2020, doi: 10.21831/jptk.v26i1.29750.

[5] Z. I. A. Karim and S. M. Maat, "Employability skills model for engineering technology students," J. Tech. Educ. Train., vol. 11, no. 2, pp. 79–87, 2019, doi: 10.30880/jtet.2019.11.02.008.

[6] R. Singh, and N. Srivastava, "Assessing the impact of student employability using decision tree classifier in education 4.0: an analysis of key factors and predictive model development," in Architecture and Technological Advancements of Education 4.0, IGI Global, 2024, pp. 178-198.

[7] N. Maneechan, and W. Jitsakul, "Generating the employment predictive model using data mining techniques," in Proc. of 2021 25th International Computer Science and Engineering Conference (ICSEC), 2021, pp. 122-127.

[8] S. Batool et al., "Educational data mining to predict students' academic performance: a survey study," Education and Information Technologies, vol. 28, no. 1, pp. 905-971, 2023.

[9] J. -A. Garcia, and J. V. Murcia, "Comparison of supervised machine learning approaches in predicting employability of students," Business and Organization Studies E-Journal, vol. 1, no. 1, pp. 121-139, 2023.

[10] R. Kannan, et al., "Predicting student's soft skills based on socio-economical factors: an educational data mining approach," JOIV: International Journal on Informatics Visualization, vo. 7, no. 3-2, pp. 2040-2047, 2023.

[11] M. H. Baffa, M. A. Miyim, and A. S. Dauda, "Machine learning for predicting students' employability," UMYU Scientifica, vol. 2, no.1, pp. 001-009, 2023.

[12] R. Haque, et al., "Classification techniques using machine learning for graduate student employability predictions," International Journal on Advanced Science, Engineering & Information Technology, vol. 14, no. 1, 2024.

[13] I. M. K. Ho, K. Y. Cheong, and A. Weldon, "predicting student satisfaction of emergency remote learning in higher education during COVID-19 using machine learning techniques," PLoS ONE, vol. 16, no. 4, pp. e0249423, 2021, doi:10.1371/journal.pone.0249423.

[14] A. Bailly et al., "Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models," Computer Methods and Programs in Biomedicine, vol. 213, 2022, doi:10.1016/j.cmpb.2021.106504.

[15] E. Y. Boateng, J. Otoo, and D. A. Abaye, "Basic tenets of classification algorithms k-nearest-neighbor, support vector machine, random forest and neural network: a review," Journal of Data Analysis and Information Processing, vol. 8, no. 4, pp. 341-357, 2020.

[16] P. Probst, M. N. Wright, and A. L. Boulesteix, "Hyperparameters and tuning strategies for random forest," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 9, no. 3, 2019, doi: 10.1002/widm.1301.