



## OPT-BAG Model for Predicting Student Employability

Minh-Thanh Vo<sup>1</sup>, Trang Nguyen<sup>2</sup> and Tuong Le<sup>3,4,\*</sup>

<sup>1</sup>Faculty of Information Technology, HUTECH University, Ho Chi Minh City, Vietnam

<sup>2</sup>Faculty of Information Technology, Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam

<sup>3</sup>Laboratory for Artificial Intelligence, Institute for Computational Science and Artificial Intelligence, Van Lang University, Ho Chi Minh City, Vietnam

<sup>4</sup>Faculty of Information Technology, School of Technology, Van Lang University, Ho Chi Minh City, Vietnam

\*Corresponding Author: Tuong Le. Email: tuong.lecung@vlu.edu.vn

Received: 23 January 2023; Accepted: 18 May 2023; Published: 30 August 2023

**Abstract:** The use of machine learning to predict student employability is important in order to analyse a student's capability to get a job. Based on the results of this type of analysis, university managers can improve the employability of their students, which can help in attracting students in the future. In addition, learners can focus on the essential skills identified through this analysis during their studies, to increase their employability. An effective method called OPT-BAG (OPTimisation of BAGging classifiers) was therefore developed to model the problem of predicting the employability of students. This model can help predict the employability of students based on their competencies and can reveal weaknesses that need to be improved. First, we analyse the relationships between several variables and the outcome variable using a correlation heatmap for a student employability dataset. Next, a standard scaler function is applied in the preprocessing module to normalise the variables in the student employability dataset. The training set is then input to our model to identify the optimal parameters for the bagging classifier using a grid search cross-validation technique. Finally, the OPT-BAG model, based on a bagging classifier with optimal parameters found in the previous step, is trained on the training dataset to predict student employability. The empirical outcomes in terms of accuracy, precision, recall, and F1 indicate that the OPT-BAG approach outperforms other cutting-edge machine learning models in terms of predicting student employability. In this study, we also analyse the factors affecting the recruitment process of employers, and find that general appearance, mental alertness, and communication skills are the most important. This indicates that educational institutions should focus on these factors during the learning process to improve student employability.

**Keywords:** Ensemble classifier; grid search cross-validation; OPT-BAG; student employability



## 1 Introduction

Machine learning is a sub-field of artificial intelligence (AI) in which computers are taught to learn from data without being explicitly programmed. The two main tasks in machine learning are supervised and unsupervised learning. Unsupervised learning [1,2] is used with unlabeled observations to cluster data (grouping similar items together) or to carry out dimensionality reduction (reducing the number of dimensions in a dataset). Supervised learning or classification is a more common type of machine learning, in which a machine learning model is trained on a labelled dataset. The goal is to use historical data to teach the model to accurately predict correct outcomes for new datasets. Supervised learning algorithms are used for tasks such as classification (e.g., determining whether or not a company is bankrupt based on financial ratios) and regression (e.g., predicting house prices in a specific area). Following the recent developments in this area, classification has been applied in numerous ways in a wide range of domains, such as economics [3,4], medical diagnosis [5,6], construction [7,8], and cybernetics [9,10]. For instance, in the area of bankruptcy prediction [3,11,12], a classification problem has been used to predict bankruptcies within a certain number of years by leveraging historical financial data. The problem of bankruptcy prediction can be considered a binary classification in which the bankruptcy dataset contains two types of companies: bankrupt and ordinary companies. Similarly, churn prediction [4,13] can also be considered as a binary classification problem in which a model detects which customers are likely to leave a service or cancel a subscription to a service. In the field of medical diagnosis, Vo et al. [5] used deep learning for multi-class classification of shoulder implant X-ray images. In addition, Le et al. [14] proposed a method for identifying self-care problems for children with a disability, based on the SCADI dataset with seven classes. More recently, Vo et al. [9] developed a novel framework for trash classification that considered three classes of images: organic, inorganic, and medical trash. This model could help enhance the capabilities of automatic intelligent waste sorter machines.

In the light of the ongoing Industry 4.0 revolution and the notable developments in AI, universities all over the world need to ensure that their graduates have suitable skills for employment their graduates. Companies are beginning to lay off large numbers of employees; for example, on 18th January 2023, Microsoft announced they would make 10,000 employees redundant as part of their broader cost-cutting efforts [15]. Almost immediately, on 20th January 2023, Google's parent company, Alphabet Inc, announced that it would also cut around 12,000 employees [16]. All of the factors described above are putting significant pressure on the educational system due to the fierce competition for employment, and the development of an advanced technique to improve the employability of students is therefore of the utmost necessity. For instance, identifying the critical factors that affect employability can be a great help to students when they are job hunting. If their weaknesses and strengths can be recognised, students and instructors can plan their careers better, and program directors can anticipate and improve their curriculum to build new competencies for educating, training, and reskilling their learners. This study focuses on employability prediction through data-driven and machine-learning methods, motivated by the need to investigate and predict emerging employability factors.

There have been many studies that have investigated data mining and machine learning techniques for the prediction of employability [17–22]. Mezhoudi et al. [17] conducted a survey of current approaches, research challenges and applications in regard to employability prediction. There are many studies in the literature on employability prediction in several different contexts in terms of the level of education (high school, Bachelor's, Master's, and PhD level), field (information technology, engineering, psychology, educational technology, etc.) and country (United States, Philippines, Malaysia, Saudi Arabia, India, and others). Despite the interesting findings in this area, existing studies have

focused on a particular region, field, and level of education, meaning that these methods cannot be applied to other datasets. The student employability dataset [23], which was collected from different university agencies in the Philippines from 2015 to 2018, is a new and interesting resource for student employability prediction. In this study, we present an effective method called OPT-BAG, based on the bagging classifier, with optimal parameters for the problem of student employability prediction based on this dataset. The primary contributions of this research are as follows.

- We first analyse the relationships between several variables and the outcome variable through the use of a correlation heatmap.
- A standard scaler function is used in the preprocessing module to normalise the variables in the student employability dataset.
- The training set is input to our model to identify the optimal parameters for the bagging classifier using a grid search cross-validation (GridSearchCV) technique.
- The bagging classifier, with the optimal parameters found in the previous step, is trained on the dataset and used to predict student employability. Experimental results indicate that our proposed approach outperforms other cutting-edge machine learning models for this prediction task.
- Using the OPT-BAG model, we analyse the factors affecting the recruitment process of employers. The results indicate that general appearance, mental alertness, and communication skills are the most important. Educational institutions should therefore focus on these factors to improve student employability.

The rest of this article is organised as follows. [Section 2](#) reviews related work on employability prediction. The student employability dataset and the proposed OPT-BAG model for predicting student employability are introduced in [Section 3](#). Empirical experiments are described in [Section 4](#). [Section 5](#) presents the conclusion of this study and suggests several directions for future work.

## 2 Related Work

Mohamed et al. [18] introduced a data mining technique for predicting employability in Morocco. The authors collected an experimental dataset through an employability survey at a Morocco university. A data mining approach was then developed for predicting employability and identifying areas for future improvement. Next, Casuat et al. [19] utilised machine learning to predict students' employability. Three learning algorithms, decision tree (DT), random forest (RF), and support vector machine (SVM), were used to predict employability. The results showed that SVM gave the best results on the student employability dataset [23]. In 2022, Wu [20] used a convolution neural network (CNN) to predict trends in the employment of college graduates based on a dataset collected from 20 comprehensive colleges and universities in China. For this dataset, 4,000 observations were performed, and 3,812 valid samples were obtained. The purpose of the study was to understand the developing trends in the unidirectional flow of university graduates from the city to the labour market. The authors also provided new ideas and directions for solving practical problems, such as the difficulty in finding employment for college students and the mode of talent training in colleges and universities in China. Later, Saidani et al. [21] utilised a gradient boosting model for predicting student employability in the context of internship, based on a private dataset collected via an online survey of the Information Systems department at the College of Computer and Information Sciences (CCIS) in Princess Nourahbint Abdulrahman University (PNU), Saudi Arabia. Finally, Raman et al. [22] discussed the role of predictive analytics in explaining the employability of a student on a Master's

in Business Management (MBA) program, using a private dataset collected from reputable business schools (B-Schools) in India. The methods described above are summarised in [Table 1](#).

**Table 1:** Related works: approaches and experimental datasets

Reference	Approach	Dataset
Mohamed et al. [18]	Data mining technique	Private dataset
Casuat et al. [19]	SVM	Student employability dataset [23]
Wu [20]	CNN	Private dataset collected in China
Saidani et al. [21]	Gradient boosting model	Private dataset collected via an online survey at CCIS in PNU, Saudi Arabia
Raman et al. [22]	Discussion of the role of predictive analytics	Private dataset collected from reputable B-Schools, India

The studies in [Table 1](#) are local to the region, field, and level of education represented in each individual dataset, and cannot be applied to our student employability dataset [23], which was collected from different university agencies in the Philippines from 2015 to 2018. We therefore propose the OPT-BAG model for the problem of student employability prediction based on this dataset.

### 3 Materials and Methods

We introduce the student employability dataset in [Section 2.1](#). The proposed model, called OPT-BAG, was developed based on a bagging classifier, and GridSearchCV [24] is used to find the optimal parameters of the bagging classifier for the student employability dataset. The bagging classifier is described in [Section 2.2.](#), and our OPT-BAG model is presented in [Section 2.3](#).

#### 3.1 Student Employability Dataset

The student employability dataset [23] was collected from different university agencies in the Philippines from 2015 to 2018, via a process that was compliant with the Data Privacy Act of the Philippines. It consists of mock job interview results, with 2,982 observations and nine features, which are shown in [Table 2](#). The first eight features, numbered F1 to F8, are general appearance, manner of speaking, physical condition, mental alertness, self-confidence, ability to present ideas, communication skills, and student performance rating, respectively. The values of these features were rated from one to five, to reflect the student's score for the corresponding ability. The last feature (F9) is student employability, which can take one of two values: employable or less employable.

**Table 2:** Variables in the student employability dataset

No.	Feature name	Feature description
1	F1	General appearance
2	F2	Manner of speaking
3	F3	Physical condition

(Continued)

**Table 2 (continued)**

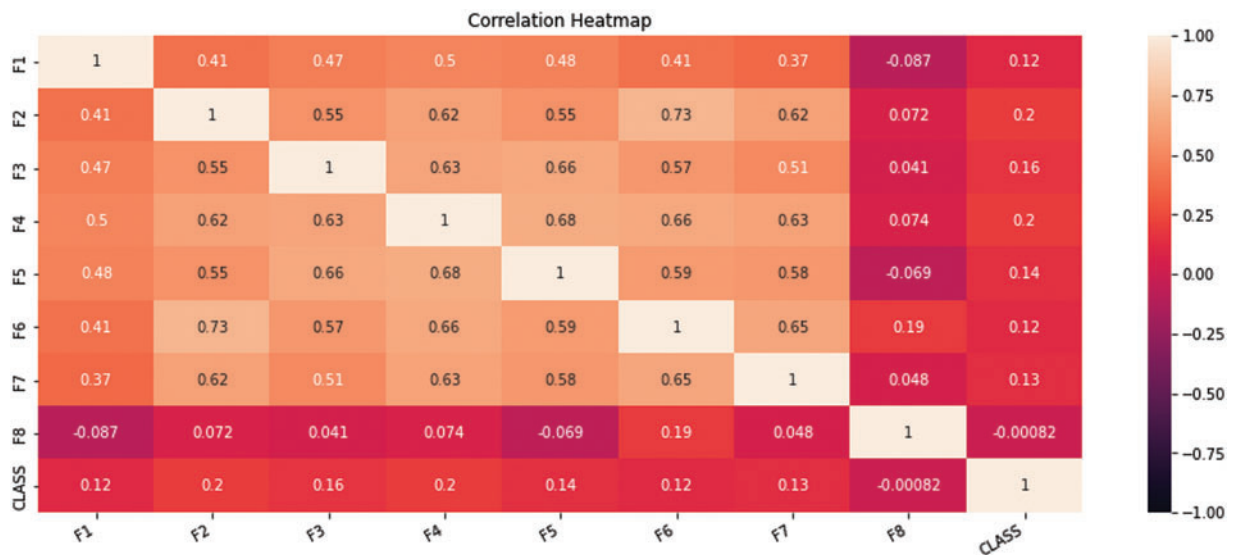
No.	Feature name	Feature description
4	F4	Mental alertness
5	F5	Self-confidence
6	F6	Ability to present ideas
7	F7	Communication skills
8	F8	Student performance rating
9	F9	Employable/less employable

We then drew up a correlation heat map of the student employability dataset. A correlation heatmap is a type of plot that visualises the strength of the relationships between pairs of variables in a dataset. The correlation coefficient between two variables  $X$  and  $Y$  can be calculated using Eq. (1) as follows:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{1}$$

where  $\bar{x}$  is the mean value of  $X$ , and  $\bar{y}$  is the mean value of  $Y$ .  $x_i$  and  $y_i$  represent the different values of  $X$  and  $Y$ . The value of the correlation coefficient can take any value in the range  $[-1, 1]$ . A positive correlation is found when two variables have the same sign, and a negative correlation occurs when the two variables have opposite signs. Correlation heatmaps are used to understand which variables are related to each other, and the strength of these relationships.

From the results in Fig. 1, we see that there are no strong positive or negative correlations between any of the features and the overall employability. The features with the highest absolute values for the correlation are manner of speaking and mental alertness, with values of 0.2.



**Figure 1:** Correlation heatmap for the student employability dataset

### 3.2 Bagging Classifier

---

#### Algorithm 1. Bagging classifier

---

**Input:** Training dataset  $D$ , Number of bootstrap samples  $T$

**Output:** Classifier  $C$

1 Let  $C =$  be the ensemble

2 For  $k = 1, \dots, T$ :

3 Take the bootstrap samples  $D_k$  from  $D$

4 Build a decision tree  $DT_k$  using  $D_k$  as the training set

5 Add  $DT_k$  to the current ensemble,  $C = C \cup DT_k$

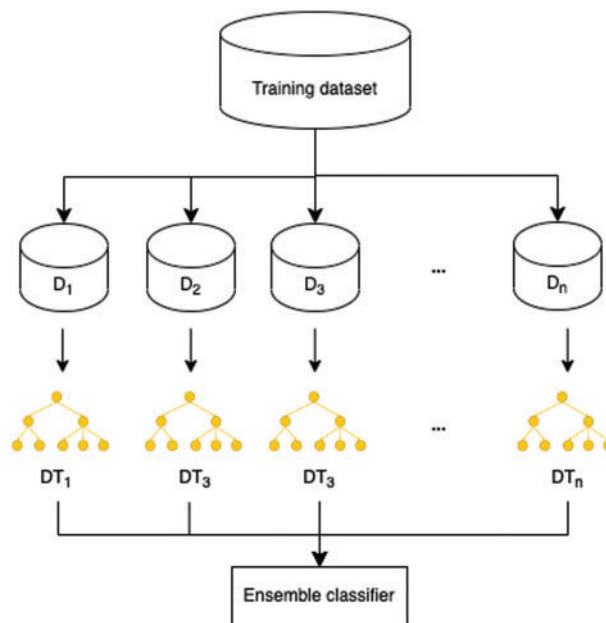
6 End for

7 Return  $C$

---

The bagging classifier [25] is an ensemble classifier, as shown in Algorithm 1. This algorithm first fits several base classifiers to random subsets from the original dataset (Lines 3 to 5), and then aggregates their individual predictions to make a final prediction.

Fig. 2 shows an overview of the bagging classifier used in the proposed model. The basic idea is to repeatedly sample observations (with replacement) from the dataset to create  $n$  bootstrap samples ( $D_1, D_2, D_3, \dots, D_n$ ) as shown in Fig. 2. The bagging algorithm uses a “bootstrap sample” to fit the different estimators ( $DT_1, DT_2, DT_3, \dots, DT_n$ ). A DT is used as an estimator in the proposed approach. The  $n$  estimators are combined in the ensemble classifier to create a final prediction.



**Figure 2:** Overview of the bagging classifier

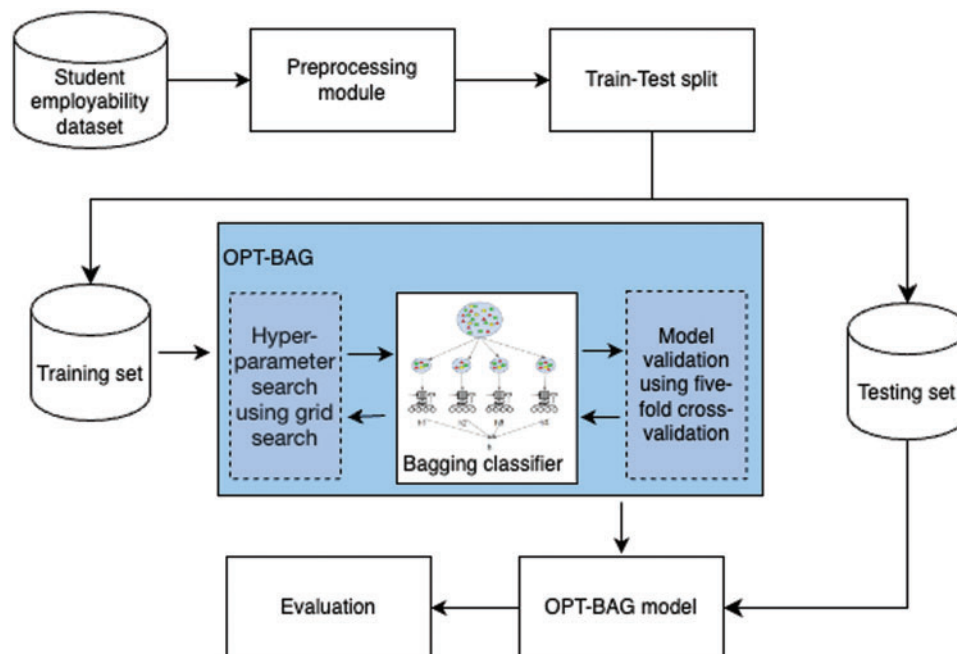
### 3.3 OPT-BAG Model

The student employability dataset is first input to the preprocessing module (see Fig. 3), which applies a standard scaler function [24] to normalise all features in this dataset. This function is a scaling technique that makes the data scale-free, by converting the statistical distribution of the data into a format with a mean at zero and a standard deviation of one. The new values are calculated using Eq. (2):

$$z = \frac{x - \mu}{\sigma} \quad (2)$$

where  $\mu$  is the mean of all values, and  $\sigma$  is the standard deviation.

After normalisation, the student employability dataset is divided into two subsets using the train-test split module (see Fig. 3), representing the training set and the testing set. The training set is input to the OPT-BAG module (shown as a blue rectangle in Fig. 3).



**Figure 3:** Flowchart for the OPT-BAG model

The proposed model uses GridSearchCV to optimise the hyperparameters of the bagging classifier. This process finds the best combination of hyperparameters to give optimal performance results based on a grid of hyperparameters. In the training phase, GridSearchCV creates many models, each with a unique combination of hyperparameters. The goal of GridSearchCV is to train each of these models and evaluate their performance using cross-validation. Finally, the model that gives the best results is selected. In this step, this model tunes the max depth of the base estimator (DT classifier), number of estimators, bootstrap, and max samples with the following sets {None, 1, 2, 3, 4, 5}, {2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 20, 30}, {True, False}, and {0.5, 0.6, 0.7, 0.8, 0.9, 1.0}, respectively.



The optimal parameters for the bagging classifier are presented in [Table 3](#). The base estimator of the bagging classifier is the DT classifier, with the number of estimators (`n_estimators`) set to two, meaning that there are two DTs in the OPT-BAG module. The value of `base_estimator__max_depth` is zero, indicating that the algorithm did not set the `max_depth` parameter in the DT classifier. In addition, the optimal values for `bootstrap` and `max_samples` are ‘false’ and 0.6, respectively.

**Table 3:** Optimal parameters for the OPT-BAG model

No.	Parameter	Value/selection
1	Estimator	DT classifier
2	<code>n_estimators</code>	2
3	<code>Base_estimator__max_depth</code>	None
4	<code>Bootstrap</code>	False
5	<code>Max_samples</code>	0.6

The OPT-BAG module returns a bagging model that has been trained with the optimal parameters (labelled the OPT-BAG model in [Fig. 3](#)). This OPT-BAG model can be used to predict employability based on the testing set. Finally, the evaluation module calculates the values of accuracy, precision, recall, and F1 metrics for the proposed approach.

## 4 Experiments

In this section, we describe some experiments that were carried out to demonstrate the effectiveness of the proposed model in terms of predicting student employability. [Section 3.1](#) explains the settings used in the experiments, such as the method of comparison, the environment used to execute these methods, and the performance metrics. [Section 3.2](#) presents the experimental results. Finally, an experiment conducted to analyse the importance of each feature is reported in [Section 3.3](#).

### 4.1 Experimental Settings

In this section, we evaluate the proposed OPT-BAG model and several other machine learning classification models, such as RF, extreme gradient boosting [26], AdaBoost, gradient boosting, DT, support vector classifier [19], and CatBoost [27], on the student employability dataset. The experimental methods were implemented using the Scikit-learn package [24] and executed on a machine with a Ubuntu operating system and an Intel Core i7-4790K (4.0 GHz × eight cores), 32 GB of RAM, and a GeForce GTX 1080 Ti. To demonstrate the effectiveness of the proposed model, we applied four standard performance metrics (accuracy, precision, recall, and F1) to the classification problem in order to compare the experimental methods. These metrics were computed using [Eqs. \(3\)–\(6\)](#), as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$



$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

The student employability dataset was first divided into two subsets, representing the training and testing sets. The training set accounted for 75%, while the testing set accounted for 25%. More specifically, the training set contained 2,236 observations, while the remaining 746 observations were allocated to the testing set.

### 4.2 Experimental Results

The confusion matrix for each experimental approach is shown in Fig. 4. The results for the RF, DT, and the CatBoost classifiers seem to be equivalent, as each gives a total of 69 observations that are wrong. RF and CatBoost both give 25 false positive observations and 44 false negative observations, while DT gives 22 false positives and 47 false negatives (Figs. 4A, 4E, and 4H). In contrast, the extreme gradient boosting, AdaBoost, gradient boosting, and support vector classifiers do not give good results, with 129, 260, 126, and 95 wrong predictions, respectively (Figs. 4B–4D, and 4G). The proposed approach achieves the best results, with only 66 wrong predictions (32 false positives and 34 false negatives, Fig. 4F).

From the confusion matrices in Fig. 4, we can calculate the values of four standard metrics for classification problems, namely the accuracy, precision, recall, and F1. Table 4 presents the experimental results for each of the experimental approaches on the student employability dataset. For the accuracy metric, our proposed approach achieved the best value of 91.15%, followed by the DT, RF, and CatBoost classifiers with 90.75%. In fifth place was the support vector classifier with 87.27%, and the last three were gradient boosting, extreme gradient boosting, and AdaBoost.

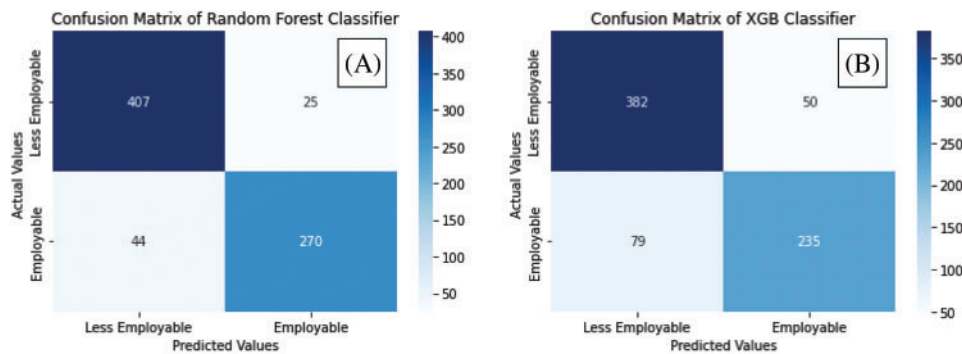
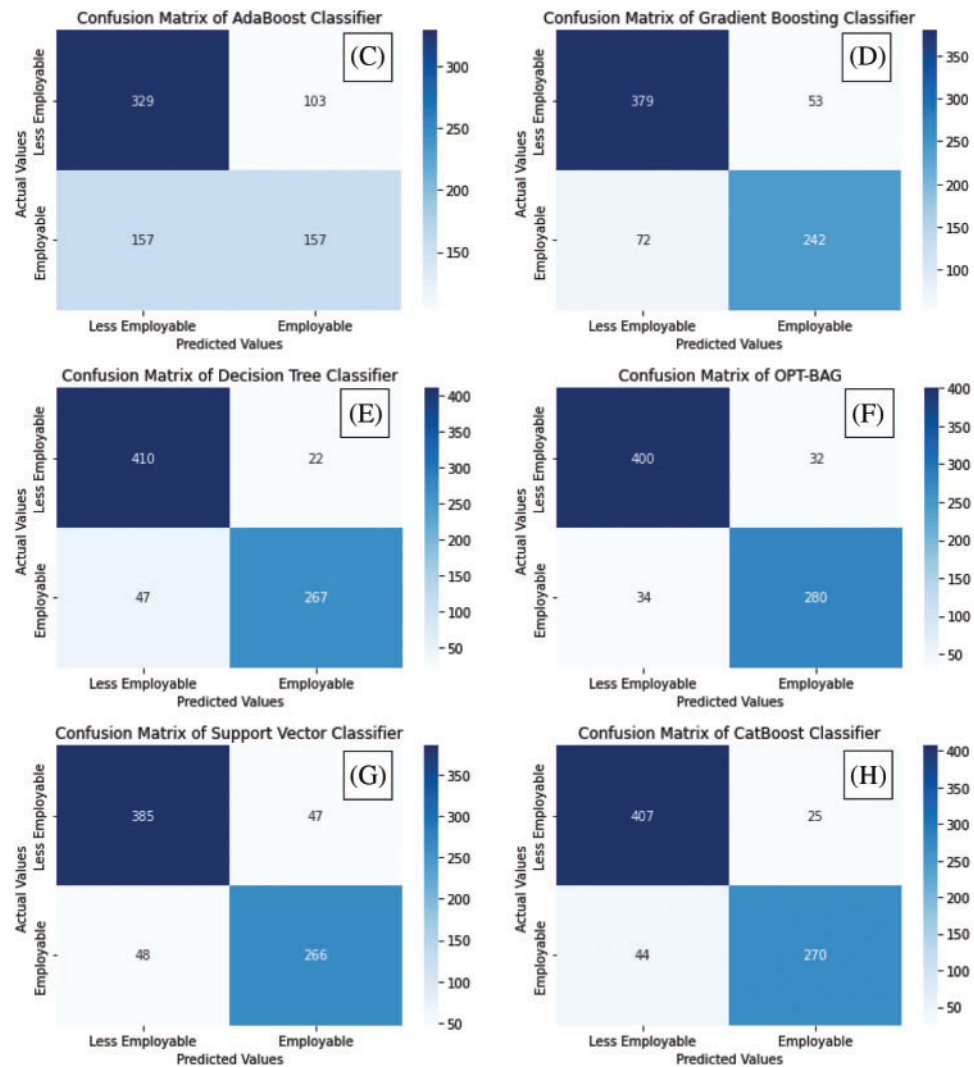


Figure 4: (Continued)



**Figure 4:** Confusion matrices for the RF, extreme gradient boosting, AdaBoost, gradient boosting, DT, OPT-BAG, support vector, and CatBoost classifiers for the student employability dataset

**Table 4:** Results for each experimental approach on the student employability dataset

No.	Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
1	RF	90.75 (2)	91.53 (2)	85.99 (2)	88.67 (2)
2	Extreme gradient boosting	82.71 (7)	82.46 (7)	74.84 (7)	78.47 (7)
3	AdaBoost	65.15 (8)	60.38 (8)	50.00 (8)	54.70 (8)
4	Gradient boosting	83.24 (6)	82.03 (6)	77.07 (6)	79.47 (6)
5	DT	90.75 (2)	<b>92.39 (1)</b>	85.03 (4)	88.56 (4)

(Continued)

**Table 4 (continued)**

No.	Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
6	OPT-BAG	<b>91.15 (1)</b>	89.74 (4)	<b>89.17 (1)</b>	<b>89.45 (1)</b>
7	Support vector classifier [19]	87.27 (5)	84.71 (5)	84.98 (5)	84.85 (5)
8	CatBoost classifier	90.75 (2)	91.53 (2)	85.99 (2)	88.67 (2)

For the precision metric, DT achieves the best result, with 92.39%, followed by RF and CatBoost with 91.53% and our proposed approach with 89.74%. The support vector classifier, extreme gradient boosting, gradient boosting and AdaBoost take the last three places. For the recall metric, our approach gave the first result, with 89.17% while RF, CatBoost and DT yielded the second, third, and fourth best results with 85.99%, 85.99%, and 85.03%, respectively. For the last metric (F1), the proposed approach also occupies the top position with the F1 value of 89.45.

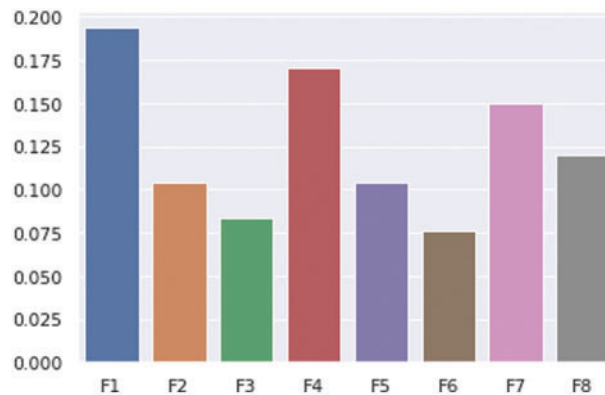
Overall, the proposed method achieves excellent and stable results based on all four-standard metrics of accuracy, precision, recall, and F1. The accuracy and F1 of the proposed method were the highest, and it struck a balance between precision and recall. We can therefore conclude that the proposed model gives the best performance for the problem of predicting employability based on the student employability dataset.

### 4.3 Relative Importance of Factors

In this experiment, we analysed the most critical factors affecting the employability of students, based on our dataset, by investigating the importance of each of the features considered in the OPT-BAG model. To do this, we calculated an importance score for each feature based on its contribution to the prediction of the target variable. Feature importance scores play an essential role in a predictive model, as they provide insight into the data and the machine learning model. In this part of the study, we aimed to understand which factors were more important in terms of employability. From the results in Fig. 5, it is easy to see that F1 (general appearance), F4 (mental alertness) and F7 (communication skills) are the three most important factors. This indicates that the general appearance, mental alertness, and communication skills of a candidate has a significant effect on recruitment by the employer. Learners therefore need to focus on improving these abilities throughout the learning process. In addition, educators need to create programs to enhance the general appearance, mental alertness, and communication skills of students in order to improve their employability.

### 4.4 Discussion

In this section, we have reported the results of two experiments involving a performance comparison with several machine learning models for classification (Section 4.2) and an analysis of the most critical factors affecting employability, based on the student employability dataset (Section 4.3). The first experiment confirmed that our OPT-BAG model achieves excellent and stable results compared with other experimental methods based on the four-standard metrics of accuracy, precision, recall, and F1. In the second experiment, we used the OPT-BAG model to analyse the most critical factors affecting employability based on the student employability dataset. The results of this experiment indicated that general appearance, mental alertness, and communication skills were the most significant factors affecting the employment process for students.



**Figure 5:** Analysis of the importance of features based on the student employability dataset

The primary limitation of this study is that the proposed model depends entirely on the data in the student employability dataset [23]. When the observations in a dataset are not sufficiently diverse, this will lead to bias problems and inaccurate predictions for new data, which means that the proposed model needs to be verified using new datasets from other regions. In addition, although our experiments identified general appearance, mental alertness, and communication skills as the most important factors, verifying these results would be a complex process, and time and in-depth studies would be needed to observe the results after these aspects had been improved.

## 5 Conclusion

This study has introduced an effective method called OPT-BAG for the problem of predicting student employability. First, the relationships between several variables and the target variable were analysed using a correlation heatmap. Next, a standard scaler function was used to normalise the variables in the student employability dataset. The training set was passed to the proposed model to identify the optimal parameters for the bagging classifier using a GridSearchCV technique. Finally, the bagging classifier was trained on the training dataset with the optimal parameters found in the previous step.

Our experiments show that the OPT-BAG model achieves excellent and stable results based on the metrics of accuracy, precision, recall, and F1. The proposed model also outperforms other cutting-edge machine learning models such as RF, extreme gradient boosting, AdaBoost, gradient boosting, DT, support vector classifier, and CatBoost in terms of precision, recall, accuracy, and F1. In addition, we analysed the main factors affecting the recruitment process by employers and found that the general appearance, mental alertness, and communication skills of a candidate was the most important feature. Educators therefore need to develop programs to allow students to improve these aspects to improve their employability.

Since the proposed OPT-BAG method is data-driven, its most important weakness is that the model depends entirely on the data in the student employability dataset. When data are not well prepared, and include noisy or incomplete items, this will mean that the proposed method works very well on the experimental dataset but does not work well on other datasets of the same type. Hence, data preparation and data collection are extremely important.

In future work, we will focus on improving the predictive performance of our model for the problem of student employability by applying several advanced data mining techniques. We will also

collect several datasets on student employability from our country to verify the proposed approach. In addition, several improvements in the undergraduate curriculum at a university will be investigated, with a focus on the factors that most influence employability, and we will evaluate their effectiveness.

**Funding Statement:** The authors received no specific funding for this study.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

- [1] Q. T. Bui, B. Vo, V. Snasel, W. Pedrycz, T. P. Hong *et al.*, “SFCM: A fuzzy clustering algorithm of extracting the shape information of data,” *IEEE Transactions on Fuzzy Systems*, vol. 29, no. 1, pp. 75–89, 2020.
- [2] Q. T. Bui, B. Vo, H. A. N. Do, N. Q. V. Hung and V. Snasel, “F-mapper: A fuzzy mapper clustering algorithm,” *Knowledge-Based Systems*, vol. 189, no. Suppl. 4, 105107, 2020.
- [3] T. Le, “A comprehensive survey of imbalanced learning methods for bankruptcy prediction,” *IET Communications*, vol. 16, no. 5, pp. 433–441, 2022.
- [4] X. Wu, P. Li, M. Zhao, Y. Liu, R. G. Crespo *et al.*, “Customer churn prediction for web browsers,” *Expert Systems with Applications*, vol. 209, no. 1, 118177, 2022.
- [5] T. N. N. Vo, H. V. Do, M. T. Tran, V. H. Pham, P. T. Nguyen *et al.*, “VNU-diagnosis: A novel medical system based on deep learning for diagnosis of periapical inflammation from X-Rays images,” *Journal of Intelligent & Fuzzy Systems*, vol. 43, no. 1, pp. 1417–1427, 2022.
- [6] M. M. Ahsan and Z. Siddique, “Machine learning-based heart disease diagnosis: A systematic literature review,” *Artificial Intelligence in Medicine*, vol. 128, no. 3, 102289, 2022.
- [7] D. K. Thai, D. N. Le, Q. H. Doan, T. H. Pham and D. N. Nguyen, “Classification models for impact damage of fiber reinforced concrete panels using Tree-based learning algorithms,” *Structures*, vol. 53, pp. 119–131, 2023.
- [8] Q. H. Doan, S. H. Mai, Q. T. Do and D. K. Thai, “A cluster-based data splitting method for small sample and class imbalance problems in impact damage classification,” *Applied Soft Computing*, vol. 120, 108628, 2022.
- [9] A. H. Vo, H. S. Le, M. T. Vo and T. Le, “A novel framework for trash classification using deep transfer learning,” *IEEE Access*, vol. 7, pp. 178631–178639, 2019.
- [10] T. K. Tran, K. T. Huynh, D. N. Le, A. Muhammad and H. M. Dinh, “A deep trash classification model on Raspberry Pi 4,” *Intelligent Automation & Soft Computing*, vol. 35, no. 2, pp. 2479–2491, 2023.
- [11] T. Le, B. Vo, H. Fujita, N. T. Nguyen and S. W. Baik, “A fast and accurate approach for bankruptcy forecasting using squared logistics loss with GPU-based extreme gradient boosting,” *Information Sciences*, vol. 494, no. 1, pp. 294–310, 2019.
- [12] T. Le, M. T. Vo, B. Vo, M. Y. Lee and S. W. Baik, “A hybrid approach using oversampling technique and cost-sensitive learning for bankruptcy prediction,” *Complexity*, 8460934, 2019. <https://doi.org/10.1155/2019/8460934>
- [13] L. Geiler, S. Affeldt and M. Nadif, “An effective strategy for churn prediction and customer profiling,” *Data & Knowledge Engineering*, vol. 142, no. 5, 102100, 2022.
- [14] T. Le and S. W. Baik, “A robust framework for self-care problem identification for children with disability,” *Symmetry*, vol. 11, no. 1, 89, 2019.
- [15] C. Thorbecke and H. Ziady, “Microsoft is laying off 10,000 employees,” 2023. [Online]. Available: <https://edition.cnn.com/2023/01/18/tech/microsoft-layoffs/index.html>
- [16] J. Dastin, “Alphabet cuts 12,000 jobs after pandemic hiring spree, refocuses on AI,” 2023. [Online]. Available: <https://www.reuters.com/business/google-parent-lay-off-12000-workers-memo-2023-01-20>

- [17] N. Mezhoudi, R. Alghamdi, R. Aljunaid, G. Krichna and D. Düşteğör, “Employability prediction: A survey of current approaches, research challenges and applications,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 3, pp. 1489–1505, 2023.
- [18] S. Mohamed and E. Abdellah, “Data mining approach for employability prediction in Morocco,” in *Proc. ESAI*, Fez, Morocco, pp. 723–729, 2019.
- [19] C. D. Casuat and E. D. Festijo, “Predicting students’ employability using machine learning approach,” in *Proc. ICETAS*, Kuala Lumpur, Malaysia, pp. 1–5, 2020.
- [20] D. Wu, “Prediction of employment index for college students by deep neural network,” *Mathematical Problems in Engineering*, vol. 2022, 3170454, 2022. <https://doi.org/10.1155/2022/3170454>
- [21] O. Saidani, L. J. Menzli, A. Ksibi, N. Alturki and A. S. Alluhaidan, “Predicting student employability through the internship context using gradient boosting models,” *IEEE Access*, vol. 10, pp. 46472–46489, 2022.
- [22] R. Raman and D. Pramod, “The role of predictive analytics to explain the employability of management graduates,” *Benchmarking: An International Journal*, vol. 29, no. 8, pp. 2378–2396, 2022.
- [23] A. Hamoutni, “Students’ employability dataset—Philippines,” 2023. [Online]. Available: <https://www.kaggle.com/datasets/anashamoutni/students-employability-dataset/>
- [24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion *et al.*, “Scikit-learn: Machine learning in python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [25] L. Breiman, “Bagging predictors,” *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.
- [26] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. KDD*, San Francisco, CA, USA, pp. 785–794, 2016.
- [27] L. O. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush and A. Gulin, “CatBoost: Unbiased boosting with categorical features,” in *Proc. NeurIPS*, Montréal, Canada, pp. 6639–6649, 2018.