Employability Prediction of Information Technology Graduates using Machine Learning Algorithms

Article *in* International Journal of Advanced Computer Science and Applications · November 2022

DOI: 10.14569/IJACSA.2022.0131043

3 authors:

Gehad Elsharkawy
Helwan University
2 PUBLICATIONS **14** CITATIONS
SEE PROFILE

Yehia K. Helmy
Helwan University
67 PUBLICATIONS **537** CITATIONS
SEE PROFILE

Engy Yehia
Helwan University
12 PUBLICATIONS **75** CITATIONS
SEE PROFILE

Employability Prediction of Information Technology Graduates using Machine Learning Algorithms

Gehad ElSharkawy, Yehia Helmy, Engy Yehia
Dept. Business Information Systems
Faculty of Commerce and Business Administration, Helwan University, Cairo, Egypt

***Abstract*—The ability to predict graduates' employability to match labor market demands is crucial for any educational institution aiming to enhance students' performance and learning process as graduates' employability is the metric of success for any higher education institution (HEI). Especially information technology (IT) graduates, due to the evolving demand for IT professionals increased in the current era. Job mismatch and unemployment remain major challenges and issues for educational institutions due to the various factors that influence graduates' employability to match labor market needs. Therefore, this paper aims to introduce a predictive model using machine learning (ML) algorithms to predict information technology graduates' employability to match the labor market demands. Five machine learning classification algorithms were applied named Decision tree (DT), Gaussian Naïve Bayes (Gaussian NB), Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM). The dataset used in this study is collected based on a survey given to IT graduates and employers. The performance of the study is evaluated in terms of accuracy, precision, recall, and f1 score. The results showed that DT achieved the highest accuracy, and the second highest accuracy was achieved by LR and SVM.**

***Keywords*—Machine learning; IT graduates; higher education; employability; labor market**

I. Introduction

Due to the dynamically changing job market and the rapid advancements in technology. The growing demand for Information Technology (IT) professionals is one of the highest demands all over the world [1]. Human capital is one of the most important economic assets of production and is considered the main pillar for raising the standard of living and developing human resources on which countries depend in strategic planning to achieve sustainable development, as human capital represents the workforce that engages in all service, production, and consumer activities in society. As a result, higher education institutions (HEIs) produce an increasing number of graduates each year. The mismatch between the higher education outputs and the labor market demands is considered one of the major threats to economic growth which causes high unemployment rate and misplacement problems among higher education graduates in Egypt. The mismatch is due to poor collaboration between the labor market and HEIs. This lack of communication results in the wrong kind of workforce, thus errors in its production are costly [2]–[4]. Thereby, to avoid this mismatch, the HEIs have to ensure the graduates' employability.

Machine learning (ML) techniques can be used to predict the employability signals of IT graduates and identify the most significant factors affecting their employability as early as possible so appropriate actions can be taken to enhance their employability in order to equip them with the appropriate knowledge and skills before they enter the dynamic job market.

There is increasing interest in applying machine learning in higher education, according to certain prior studies to predict the graduates' employability but still, the use of automated machine learning to predict students' employability in its initial stage, ML is a subset of artificial intelligence (AI) in which computers analyze large datasets to learn patterns that will make predictions for new data, in contrast to traditional computer methodologies. In traditional reasoning, algorithms are a set of explicitly defined instructions that computers use to describe or solve problems [5], [6]. As a result, in the hiring process, graduates with experience are in high demand due to high productivity and low training cost than those who did not have any experience. HEIs must undergo frequent evaluations to provide future IT graduates with the demanded skills as it is considered the main factor to produce this workforce [7].

The earlier studies have shown a great interest in examining the mismatch between HEIs output and labor market demands. By applying different ML algorithms. However, these studies focused on one or a few features only. As a result, the two main research questions of this study are:

RQ1) What are the most significant features that affect graduates' competitive advantage to match labor market demands?

RQ2) what are the best machine learning algorithms for employability prediction of IT graduates?

The objective of this study is to develop a prediction ML model for graduates' employability status (predict whether the IT graduate is most likely to be qualified or not qualified to match labor market demands), and for better utilization of the collected dataset which can greatly help understand the extent to which IT graduates were prepared for the highly technical IT careers to enter the workforce.

The findings of this study will help:

- Reduce the gap between labor market demands and HEIs.

- Improve the IT graduates' qualifications to match labor market demands.
- Provide valuable insights for guiding HEIs to make better long-term plans for producing graduates who are knowledgeable and skilled through prediction of their employability status.
- Contribute significantly to the placement process for employers.
- Decrease the high unemployment rate of IT graduates.

The rest of this paper is organized as follows: Section II presents various relevant works in the field of employability prediction. Section III describes the proposed methodology in detail. Section IV shows the results of the used algorithms and the discussion of the analysis of the used features. Section V presents the conclusions of this study with some limitations and improvements.

II. Related Work

In recent years, many researchers attempted to use machine learning in higher education to enhance graduate's features and curricula to support employability [8]. To discuss the contribution of ML in continuous quality improvement. We focused on some of the previous works that used different machine learning techniques such as Artificial Neural Network (ANN), Decision Tree (DT), K-Nearest Neighbor (KNN), Gaussian Naïve Bayes (Gaussian NB), Logistic Regression (LR), Neural Network (NN), Random Forest (RF), Naïve Bayes (NB) and Support Vector Machine (SVM).

In [9], the author predicted which students are most likely to get work after graduation by using data analytics and machine learning techniques such as SVM, LR, ANN, DT, and discriminant analysis. Also, the features used are hard skills, demographics features, extra/co-curricular activities, and internships the data were obtained from student surveys and institutional databases. The SVM classification algorithm achieved an accuracy of 87.26%.

The authors in [10] aimed to identify the most significant factors affecting graduate employability by using three classification algorithms DT, ANN, and SVM. The features used in this research are hard skills, soft skills, demographic features, extra/co-curricular activities, university features, and internships the research data were collected from institutional databases. The SVM algorithm shows 66.096% accuracy.

A web-based application is developed by [11] through applied machine learning algorithms DT, NB, and NN to predict the sustainability of IT students' skills for recruitment mainly hard skills and soft skills, the collected data were from student and recruiter surveys, the NB achieved the highest accuracy of 69%. In another research [12], supervised machine learning techniques such as LR, DT, RF, KNN, and SVM were used to predict high school students' employability for part-time jobs with local businesses the hard skills, demographic features, and extra/co-curricular activities features were used and collected from student surveys. The LR algorithm achieved an accuracy of 93%.

The authors in [13] analyzed the data from education institutions to predict the students' employability and determine the factors affecting their employability by using hard skills, soft skills, demographic features, extra/co-curricular activities, and university features then applied four ML algorithms which are DT, Gaussian NB, SVM, and KNN. The results achieved an accuracy of 98% by DT and SVM.

Furthermore, a student employability prediction system was developed by [14] using SVM, DT, RF, KNN, and LR algorithms to predict the students' employability, Institutional databases were obtained, and the hard skills, soft skills, and demographic features were used. The results of this research achieved an accuracy of 91% by the SVM algorithm. In [14], the authors identified the most predictive attributes through hard skills, soft skills, and demographic features to determine why students are most likely to get employed using graduates surveys and institutional databases, the applied and compared three methods are SVM, RF, and DT. The SVM achieved the highest accuracy of 91.22%.

The authors in [15] investigated the impact of various institution features on graduate employability using the hyperbox-based machine learning model which achieved 78% accuracy. A hybrid model was proposed by [16] for student employability prediction through a deep belief network and Softmax regression (DBN-SR) the dataset obtained from student surveys and the hard skills, soft skills, demographic features, and university features were used as the adopted features the results achieved high accuracy with 98%.

In [17] predicted the students' employability based on technical skills the institution databases were collected and the following algorithms were applied SVM, LR, DT, RF, AdaBoost, and NB, the highest accuracy achieved is 70% by the RF algorithm. Finally, the authors in [18] developed a model using various machine learning methods DT, RF, NN, and Gaussian NB to forecast candidate hiring by employing different statistical measures on feature selection such as hard skills, demographic features, and professional experience, the highest accuracy was achieved by Gaussian NB with 99%. Table I depicts and summarizes the relevant studies according to their adopted features, dataset sources, ML models, output features, and accuracy of the best-adapted model to answer RQ1.

TABLE I. COMPARISON OF RELATED STUDIES

Reference	Year	Adopted features categories	Dataset sources	ML model	Output features	Accuracy
Hugo [9]	2018	Hard skills Demographics features Extra/co-curricular activities Internship	Student surveys and Institution databases.	-SVM -ANN -LR -Discriminant analysis -DT	Employability: {Employed, Not Employed}	SVM 87.26%
Othman et al. [10]	2018	Hard skills Soft skills Demographic features Extra/co-curricular activities University features Internship	Institutional databases	-DT -ANN -SVM	Employability: {Employed, Not employed}	SVM 66.0967%
Alghamlas and Alabduljabbar [11]	2018	Hard skills Soft skills.	Student surveys and Recruiter surveys.	-DT -NB -NN	Matching to industry-required skills	Naïve Bayes 69%
Dubey and Mani [12]	2019	Hard skill Demographic features Extra/co-curricular activities.	Student surveys.	-LR -DT -RF -KNN -SVM	Hiring: {Hired, Not hired}	LR 93%
Kumar and Babu [13]	2019	Hard skills Soft skills Demographic features Extra/co-curricular activities University features.	Student surveys.	-DT -Gaussian NB -SVM -KNN.	Getting a job: {Yes, no}	DT & SVM 98%
Casuat [21]	2020	Hard skills Soft skills Demographic features	Institution databases.	-DT -RF -SVM -KNN -LR	Employability: {Employed, Less Employed}	SVM 91%
Casuat & Festijo [14]	2020	Hard skills Soft skills Demographic features.	Graduate surveys and Institution databases	-SVM -RF -DT	Employability: {Employed, Less Employed}	SVM 91.22%
Aviso et al.[15]	2020	University features.	Institution databases.	Rule-based Hyperbox model.	Employability: {Yes, no}	78%
Bai and Hira [16]	2021	Hard skills Soft skills Demographic features University features.	Student surveys.	-Softmax regression.	Employability: {Employed, Unemployed}	98%
Laddha et al. [17]	2021	Hard skills.	Institution databases.	-SVM -LR -DT -RF -AdaBoost –NB.	Placement: {Placed, Not placed}	RF 70%
Reddy et al. [18]	2021	Hard skills Demographic features Professional experience.	Employee surveys.	-DT -RF -NN -Gaussian NB.	Recruitment: {Join, Not join}	Gaussian NB 99%

III. METHODOLOGY

In this section, we will discuss the methodology of our study, the machine learning algorithms applied, and the evaluation metrics used in this study. Fig. 1 highlights the research methodology: i) Data collection; ii) applying data preprocessing; iii) Splitting the dataset into two sets, a train set to train the model and a test set to evaluate the model; iv) building our model by applying five ML classification algorithms; v) evaluating the model; vi) outcome the proposed model to predict the qualified IT graduate to meet labor market demands. To answer RQ1: What are the most significant features that affect graduates' competitive advantage to match labor market demands? we followed the methodology steps as shown below.

A. Data Source

The dataset used in this research was obtained based on a survey given to IT graduates and employers in Egypt. We created an online survey with pertinent questions and then distribute it to IT graduates including (Computers & Artificial intelligence, Business information systems, Software Engineering, and Management information systems) and several IT companies from different sectors to get the desired findings. A brief description of each feature selected, and its value is described in Table II. We classified them into four categories (Trainings, Soft skills, Hard skills, and In-demand skills) each category has the most-related features, and the values (0,1) of the first three categories indicated that "0" means the graduate does not been trained or given a specific course during their study years in the college. While "1" means the graduate has been trained or given a specific course in those skills. In the fourth category, the value (0-7) means how many courses or trainings the graduate received from those fields to be qualified for the industry requirements.

B. Data Preprocessing

Data preparation is a critical stage while creating a machine learning model as it is difficult for a machine to read the raw datasets to produce the expected results [19]. So, data preprocessing make data suitable for a machine learning model. First, we eliminate noise, missing values and make the data consistent. Then, we apply feature selection to identify the relevant features to allow classifiers to reach the optimal performance which has a greater impact on IT graduates' employability to match the labor market demands. Finally, we Split the dataset into two sets (80%) for training to train the model and (20%) for testing to test the accuracy of the model and enhance the performance of our machine learning models.

C. Prediction Models

Five different binary classification algorithms are used to predict the IT graduates' employability using the collected dataset. Because it categorizes new observations into one of two classes. The binary class in our dataset has two values (0) for a not qualified graduate that does not match labor market demands, and (1) for a qualified graduate. The number of records used in this study is 296. We used the following libraries Scikit Learn, Pandas, NumPy, Matplotlib, and Seaborn of the Python programming language. The five classification algorithms are:

Decision Tree Algorithm: is a supervised learning technique equivalent to a series of IF-THEN statements built a structure of branches and nodes based on the evidence obtained for each feature during the method learning process [10]. DT algorithm generates decision trees from training data to solve classification and regression problems. In our proposed model, the Gini method was used to create split points by finding a decision rule that produces the greatest decrease in impurity at a node.

$$G(t)=1-\sum_{i=1}^{c} Pi^2 \quad (1)$$

where $G(t)$ is the Gini impurity at node t and p_i is the proportion of observations of class c at node t. Recursively, this decision-making process is carried out until all leaf nodes are pure or a certain cutoff is achieved.

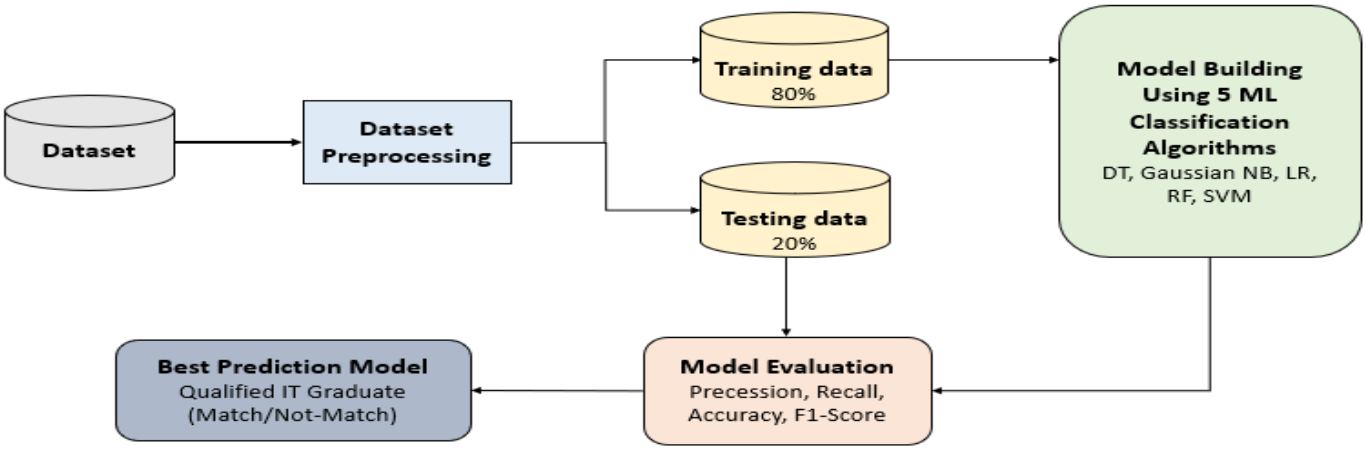

Fig. 1. The Research Methodology.

TABLE II. DESCRIPTION OF DATA FEATURES

Category	Feature	Values	Description
Trainings	Internship	(0,1)	A professional learning experience that provides meaningful work experience related to a student's field of study or career interest for a limited period of time.
	Summer training		*A period spent in a reputable company to gain relevant skills and experience in a particular field is usually conducted during July and August of each year.*
	Workshops		*A period of discussion in which people work on a particular subject by discussing it or doing activities relating to it.*
	Co-curricular activities		*The activities and learning experiences that take place in the university along with the academic curriculum by students to enhance their skills.*
Soft skills	Problem-solving	(0,1)	The act of defining a problem; finding the cause of the problem; identifying, prioritizing, and selecting alternatives for a solution; and implementing a solution.
	Creative thinking		*The ability to generate new solutions to problems.*
	Time management		The process of planning and organizing how much time to spend on specific activities.
	English proficiency		*The ability to use and understand spoken and written English.*
Hard skills	Data security	*(0,1)*	*The practice of protecting digital information.*
	Network security		The practice of protecting networks and data.
In-demand skills	Data Analytics	(0-7)	The student's knowledge and experience gained in those fields are based on their years of studies at the university through curricula and practical applications of them.
	Artificial Intelligence (AI)		
	Internet of Things (IoT)		
	Machine Learning (ML)		
	Cybersecurity		
	Data Science		
	Cloud Computing		

Gaussian Naïve Bayes (Gaussian NB) algorithm: is a variant of Naive Bayes it is a probabilistic machine learning algorithm used for many classification functions and is based on the Bayes theorem and has a strong assumption that predictors should be independent of each other [13]. The likelihood of the features in our proposed model is assumed to be Gaussian:

$$P(x_i|y)=\frac{1}{\sqrt{2\pi\sigma_y^2}}\exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \quad (2)$$

Where the parameters σ_y and μ_y are estimated using maximum likelihood.

1) Random forest algorithm: is a supervised learning algorithm. It can be used both for classification and regression. This model first generates a forest of random trees. The aim of voting to merge random trees in a forest is to eliminate the most predicted tree. If a dataset contains x features, it first chooses a random feature known as y. The algorithm then attempts to merge trees based on the expected outcome and voting procedure [20]. We used the Gini method as mentioned in (1).

2) Logistic Regression (LR) algorithm: A LR uses regression analysis, in this method a class variable that is binary classified is required for the logistic regression model [17]. Similarly, the target column named the employability class in this dataset holds two types of binary numbers "0" for a not-qualified IT graduate who has no chance of being employable to meet labor market demands, and "1" for the IT graduate who has been predicted to be qualified and match labor market requirements. In our proposed model, a linear model is included in a logistic function as follows:

$$P(y_i=1|X)=\frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \quad (3)$$

where P (y_i = 1 | X) is the probability of the *i*th observation's target value, y_i, being class 1, X is the training data, β0 and β1 are the parameters to be learned, and e is Euler's number. The logistic function's goal is to interpret its output as a probability by limiting its value to a range between 0 and 1.

3) Support Vector Machine algorithm (SVM): in SVM the classes in the dataset should be pre-defined in this model. It works by using predefined classes to classify the objects in the given dataset. It categorizes transactions by allocating one or more classes in order to increase performance accuracy [21]. We used the linear SVC (Linear Support Vector Classification).

D. Model Evaluation

To evaluate the model effectiveness, a confusion matrix with true positive (TP), false positive (FP), true negative (TN), and false negative (FN) for predicted data is formed. The performance of the study is measured with respect to the accuracy, precision, recall, and F1 score. A brief description of each is described below:

Accuracy: It is a common metric for evaluating classifier performance. It computes the ratio of correctly classified instances to the total number of instances [8]. Its formula is as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

Precision: is the ratio of true positive instances divided by the total number of instances predicted as positive [22].

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

Recall: is given as the ratio of relevant instances that are retrieved [22].

$$\text{Recall} = \frac{TP}{TP+FN} \quad (6)$$

F1 score: it is the combination of both precision and recall used to get the average value of them [20].

$$\text{F1 score} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (7)$$

IV. Results and Discussion

After data pre-processing, according to the methodology used, out of the total 296 graduates collected, 80% of the data was used as a training dataset, and 20% was kept as a test dataset. The findings related to this study are presented as follows. Fig. 2 shows the correlation matrix for the used features.

The distribution of the employability class (qualified and not qualified) graduates used in this study is illustrated in Fig. 3 the value 0 represents the number of not qualified graduates while 1 represents the number of qualified graduates. From the figure, it may be shown that most involved samples are "not qualified" graduates (82%) than the "qualified" graduates (18%).

In Fig. 4, we present the participants' distribution in terms of the features that represent the trainings taken during the graduates' years of study. According to the internship, a total of 11 participants were trained and qualified. Furthermore, 15 participants referred to this training although they were not qualified. A total of 12 participants were trained and qualified because of the summer training. Moreover, the 105 participants referred to this training even though they were not qualified. According to the co-curricular activities, a total of 41 participants were trained and qualified. Whereas 168 participants referred to this training given the fact that they were not qualified. Lastly, 37 people were trained and qualified during the workshops. And 82 participants referred to this training despite the reality that they were not qualified.

Fig. 5 illustrates the participants' distribution in terms of the features that represent the soft skills the graduates have been trained on during their years of study. As stated by the problem-solving skills a total of 39 participants were trained and qualified. As well as a number of 110 participants referred to this skill although they were not qualified. Referring to creative thinking skills, a total of 33 participants were trained and qualified. Also, a number of 74 participants referred to this skill although they were not qualified. Based on time management skills, a total of 37 participants were trained and qualified. Moreover, a number of 107 participants referred to this skill although they were not qualified. According to English proficiency skills, a total of 43 participants were trained and qualified. In addition, a number of 102 participants referred to this skill although they were not qualified.

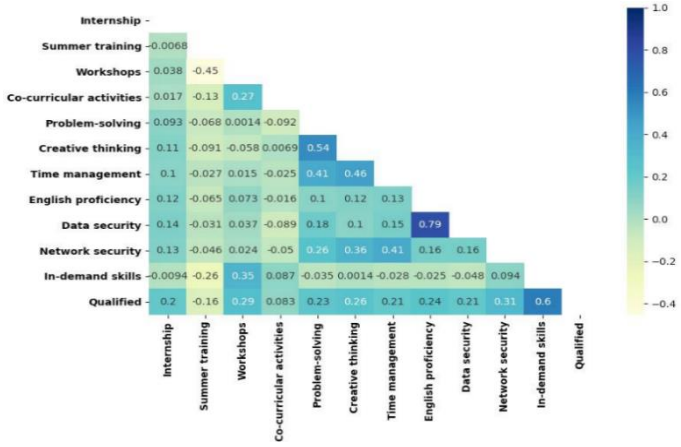

Fig. 2. Correlation Matrix of Selected Features.

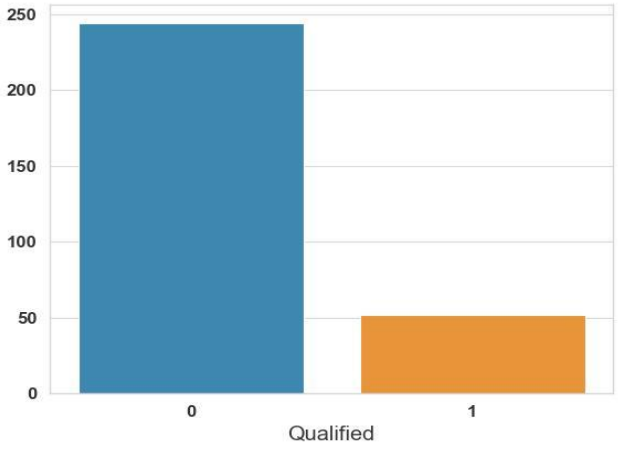

Fig. 3. Count of Employability Class (Qualified/Not Qualified).

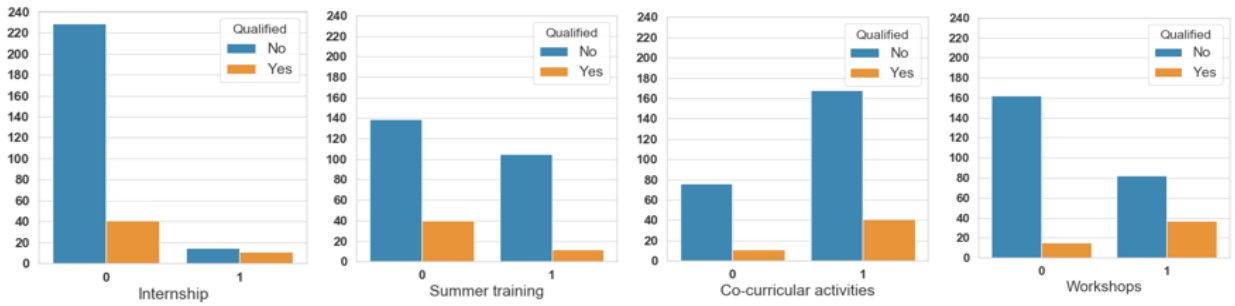

Fig. 4. Respondents' Distribution in Terms of Trainings.

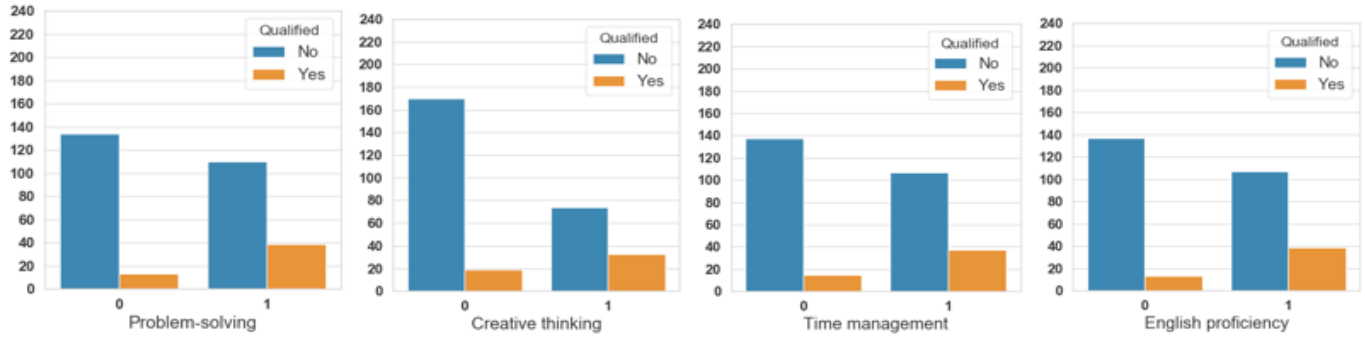

Fig. 5. Respondents' Distribution in Terms of Soft Skills.

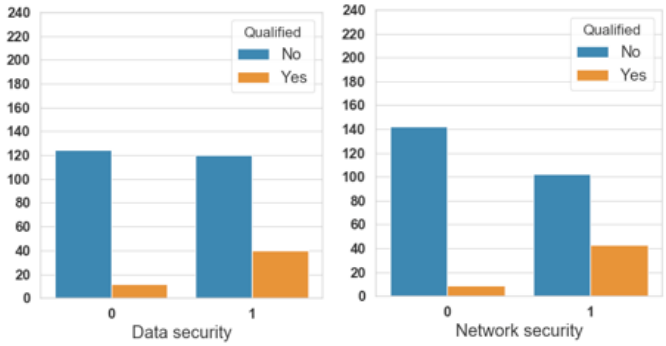

Fig. 6. Respondents' Distribution in Terms of Hard Skills.

Fig. 6 depicts the participants' distribution in terms of the features that represent the hard skills the graduates have been trained on during their years of study. According to the data security skills, a total of 39 participants were trained and qualified. Furthermore, a number of 107 participants referred to this skill although they were not qualified. A total of 40 participants were trained and qualified in network security skills, whereas a number of 120 participants referred to this skill although they were not qualified.

Fig. 7 demonstrates the participants' distribution in terms of the features that represent the in-demand skills required by the industry from the employers' perspectives of the graduates who have been trained on or given a specific course during their years of study. The 0 value means a total of 2 participants did not take any of those skills and were qualified to match labor market requirements whereas 114 participants did not take any of them and found themselves not qualified to be employable. Based on value 7, a total of 7 participants took the seven demanded skills, and they were qualified. Therefore, there are no participants who took those seven skills who were not qualified.

We applied five machine learning classification algorithms for predicting IT graduates' employability. The confusion matrix for each model is illustrated in Table III. Fig. 8 shows the outcome prediction.

Table III reveals that the DT model predicts the highest number of true positives (52 out of 59 test samples) among the five models. Furthermore, LR and SVM models predict the highest number of true negatives (13 among 59 test samples). The lowest number of false positives (0 out of 59 samples) is achieved by DT, RF, and SVM, respectively. The DT, Gaussian NB, and LR obtained the lowest number of false negatives (0 among 59).

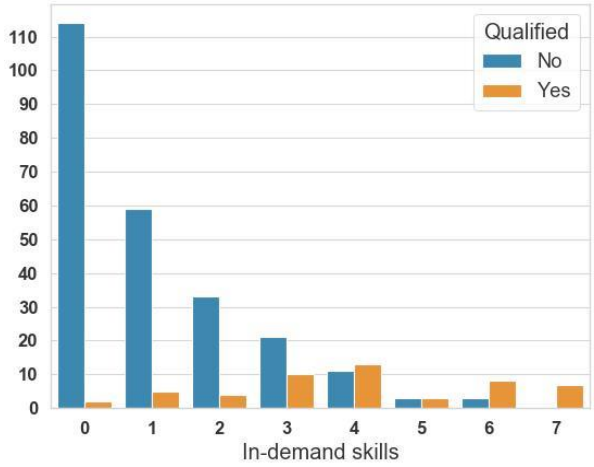

Fig. 7. Respondents' Distribution in Terms of in-demand Skills.

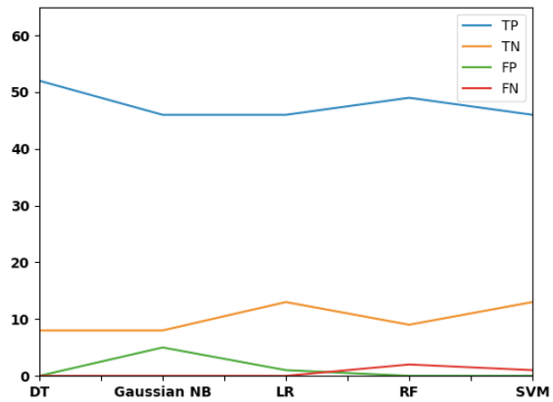

Fig. 8. Confusion Matrix for the Five Machine Learning Models.

TABLE III. CONFUSION MATRIX FOR THE MACHINE LEARNING CLASSIFICATION MODELS

	DT	Gaussian NB	LR	RF	SVM
TP	52	46	46	49	46
TN	8	9	13	9	13
FP	0	5	1	0	0
FN	0	0	0	2	1

The model performance for the employability target class in the form of a confusion matrix is presented in Table IV. In this table, the Match Class "1" means the graduates have chances of being employable and matching the labor market demands. On the other side, Not Match Class "0" denotes the graduates having no chance of being employable, the values of the row illustrating the prediction computed for both classes. As a result, the class precision, recall, and f1 score values are computed and displayed in the table. The class recall and precision values can be used to determine the classifier's overall accuracy. According to the table values, the DT classifier has the highest precision and recall, while the Gaussian NB classifier has the lowest.

The performance of the study was evaluated in terms of accuracy, precision, recall, and F1 score. The calculated performance measures are shown in Fig. 9 and Table V.

RQ2: what are the best machine learning algorithms for employability prediction of IT graduates?

Fig. 9 and Table V indicate that DT outperformed all other machine learning algorithms with a maximum accuracy of 100%, while LR and SVM achieved the second highest accuracy of 98%. DT outperformed by precision, recall, and F1 score of 100%. The second highest F1 score is achieved by LR and SVM at 98%. The second highest precision is achieved by SVM, and the second highest recall is achieved by LR. Most of the techniques have an F1 score higher than 93%, which is comparatively better.

TABLE IV. EVALUATION OF EMPLOYABILITY CLASS (QUALIFIED / NOT QUALIFIED)

Decision Tree Algorithm			
	Precision	**Recall**	**F1 score**
Match (1)	1	1	1
Not Match (0)	1	1	1
Gaussian Naive Bayes Algorithm			
	Precision	**Recall**	**F1 score**
Match (1)	0.64	1	0.78
Not Match (0)	1	0.9	0.95
Logistic Regression Algorithm			
	Precision	**Recall**	**F1 score**
Match (1)	0.93	1	0.96
Not Match (0)	1	0.98	0.99
Random Forest Algorithm			
	Precision	**Recall**	**F1 score**
Match (1)	1	0.82	0.9
Not Match (0)	0.96	1	0.98
Support Vector Machine Algorithm			
	Precision	**Recall**	**F1 score**
Match (1)	1	0.93	0.96
Not Match (0)	0.98	1	0.99

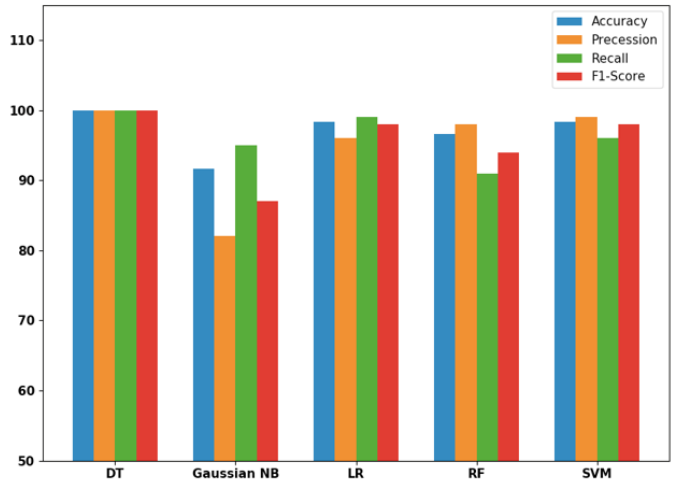

Fig. 9. Performance Measurement using Five Machine Learning Algorithms.

TABLE V. PERFORMANCE EVALUATION OF THE FIVE MACHINE LEARNING ALGORITHMS

	DT	**Gaussian NB**	**LR**	**RF**	**SVM**
Accuracy	1	0.92	0.98	0.97	0.98
Precision	1	0.82	0.96	0.98	0.99
Recall	1	0.95	0.99	0.91	0.96
F1 score	1	0.87	0.98	0.94	0.98

V. CONCLUSION

The number of information technology graduates produced by higher education institutions has been increasing every year. To overcome their unemployment situation and the mismatch between HEIs outputs and the labor market demands, there is a need for a model that can predict IT graduates' employability to match labor market requirements using machine learning techniques. Therefore, this paper proposed, discussed, and implemented five machine learning classification algorithms namely DT, Gaussian NB, LR, RF, and SVM.

This study achieved high accuracy than earlier works. The highest accuracy is achieved by DT with 100% and the second highest accuracy is achieved by LR and SVM with 98%, whereas the lowest accuracy with 92% achieved by Gaussian NB. The small size of the dataset is the main limitation of this study. From the study, we can conclude that machine learning techniques can predict IT graduates' employability with high accuracy.

The proposed model can be useful and helpful for higher education institutions to make better long-term plans for producing graduates who are knowledgeable, skilled, and fulfill the labor market needs. The findings of the features analysis indicated that moderating the curriculum to include the demanded skills required by industry and improving the teaching and learning methods by offering more training that would produce quality graduates in the following years. Also, the proposed model will be helpful for employers to contribute significantly to the placement process.

For further research, the size of the used dataset can be expanded, and various ML algorithms can be used to get better performance.

REFERENCES

[1] H. B. Kenayathulla, N. A. Ahmad, and A. R. Idris, "Gaps between competence and importance of employability skills: evidence from Malaysia," Higher Education Evaluation and Development, vol. 13, no. 2. pp. 97–112, 2019. doi: 10.1108/heed-08-2019-0039.

[2] F. Biagi, J. Castaño Muñoz, and G. Di Pietro, "Mismatch Between Demand and Supply Among Higher Education Graduates in the EU," JRC Tech. Rep., pp. 1–21, 2020, doi: 10.2760/003134.

[3] R. Assaad, C. Krafft, and D. Salehi-Isfahani, "Does the type of higher education affect labor market outcomes? Evidence from Egypt and Jordan," High. Educ., vol. 75, no. 6, pp. 945–995, Jun. 2018, doi: 10.1007/s10734-017-0179-0.

[4] M. I. Hossain, K. S. A. Yagamaran, T. Afrin, N. Limon, M. Nasiruzzaman, and A. M. Karim, "Factors Influencing Unemployment among Fresh Graduates: A Case Study in Klang Valley, Malaysia," Int. J. Acad. Res. Bus. Soc. Sci., vol. 8, no. 9, Oct. 2018, doi: 10.6007/IJARBSS/v8-i9/4859.

[5] H. Pallathadka et al., "Materials Today : Proceedings Investigating the impact of artificial intelligence in education sector by predicting student performance," Mater. Today Proc., vol. 51, pp. 2264–2267, 2022, doi: 10.1016/j.matpr.2021.11.395.

[6] H. Zeineddine, U. Braendle, and A. Farah, "Enhancing prediction of student success: Automated machine learning approach," Comput. Electr. Eng., vol. 89, 2021, doi: 10.1016/j.compeleceng.2020.106903.

[7] M. E. Oswald-Egg and U. Renold, "No experience, no employment: The effect of vocational education and training work experience on labour market outcomes after higher education," Econ. Educ. Rev., vol. 80, 2021, doi: 10.1016/j.econedurev.2020.102065.

[8] 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, doi: 10.1109/ACCESS.2022.3170421.

[9] L. S. Hugo, "Predicting Employment Through Machine Learning." https://www.naceweb.org/career-development/trends-and-predictions/predicting-employment-through-machine-learning/ (accessed Oct. 18, 2022).

[10] Z. Othman, S. W. Shan, I. Yusoff, and C. P. Kee, "Classification Techniques for Predicting Graduate Employability," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 8, no. 4–2, p. 1712, Sep. 2018, doi: 10.18517/ijaseit.8.4-2.6832.

[11] M. Alghamlas and R. Alabduljabbar, "Predicting the Suitability of IT Students' Skills for the Recruitment in Saudi Labor Market," in 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), May 2019, pp. 1–5. doi: 10.1109/CAIS.2019.8769577.

[12] A. Dubey and M. Mani, "Using Machine Learning to Predict High School Student Employability – A Case Study," in 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Oct. 2019, pp. 604–605. doi: 10.1109/DSAA.2019.00078.

[13] M. S. Kumar and G. P. Babu, "Comparative Study of Various Supervised Machine Learning Algorithms for an Early Effective Prediction of the Employability of Students," J. Eng. Sci., vol. 10, no. 10, pp. 240–251, 2019.

[14] C. D. Casuat and E. D. Festijo, "Identifying the Most Predictive Attributes among Employability Signals of Undergraduate Students," Proc. - 2020 16th IEEE Int. Colloq. Signal Process. its Appl. CSPA 2020, no. May, pp. 203–206, 2020, doi: 10.1109/CSPA48992.2020.9068681.

[15] K. B. Aviso, J. I. B. Janairo, R. I. G. Lucas, M. A. B. Promentilla, D. E. C. Yu, and R. R. Tan, "Predicting higher education outcomes with hyperbox machine learning: what factors influence graduate employability?," Chem. Eng. Trans., vol. 81, no. 2019, pp. 679–684, 2020, doi: 10.3303/CET2081114.

[16] A. Bai and S. Hira, "An intelligent hybrid deep belief network model for predicting students employability," Soft Comput., vol. 25, no. 14, pp. 9241–9254, Jul. 2021, doi: 10.1007/s00500-021-05850-x.

[17] M. D. Laddha, V. T. Lokare, A. W. Kiwelekar, and L. D. Netak, "Performance Analysis of the Impact of Technical Skills on Employability," Int. J. Performability Eng., vol. 17, no. 4, p. 371, 2021, doi: 10.23940/ijpe.21.04.p5.371378.

[18] D. Jagan Mohan Reddy, S. Regella, and S. R. Seelam, "Recruitment Prediction using Machine Learning," in 2020 5th International Conference on Computing, Communication and Security (ICCCS), Oct. 2020, pp. 1–4. doi: 10.1109/ICCCS49678.2020.9276955.

[19] S. R. Rahman, M. A. Islam, P. P. Akash, M. Parvin, N. N. Moon, and F. N. Nur, "Effects of co-curricular activities on student's academic performance by machine learning," Curr. Res. Behav. Sci., vol. 2, no. May, p. 100057, 2021, doi: 10.1016/j.crbeha.2021.100057.

[20] A. Alhassan, B. Zafar, and A. Mueen, "Predict Students' Academic Performance based on their Assessment Grades and Online Activity Data," Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 4, 2020, doi: 10.14569/IJACSA.2020.0110425.

[21] C. D. Casuat, "Predicting Students' Employability using Support Vector Machine: A SMOTE-Optimized Machine Learning System," Int. J. Emerg. Trends Eng. Res., vol. 8, no. 5, pp. 2101–2106, May 2020, doi: 10.30534/ijeter/2020/102852020.

[22] P. Thakar, P. Dr., and D. Manisha, "Role of Secondary Attributes to Boost the Prediction Accuracy of Students' Employability Via Data Mining," Int. J. Adv. Comput. Sci. Appl., vol. 6, no. 11, 2015, doi: 10.14569/IJACSA.2015.061112.