Controlling Employability Issues of Computing Science Graduates through Machine Learning-Based Detection and Identification

Wael G. Alheadary
College of Computer Science and Engineering, Taibah University, Saudi Arabia
wheadary@taibahu.edu.sa (corresponding author)

Received: 29 March 2023 | Revised: 18 April 2023 and 23 April 2023 | Accepted: 24 April 2023

ABSTRACT
The unemployment rate of graduate students in the area of computing is tremendously growing. One of the main reasons is the difference between the acquired skills from universities and the skills required from industry which is looking for potential graduates who can work in the digitally transforming framework of today’s society. Many studies have been conducted to emphasize the issue of unemployment utilizing traditional approaches. However, these methods are time-consuming and difficult to bring into effect, while involving a lot of effort, which had no definite influence or impact on the studies to date. Hence, this study proposes a predictive artificial intelligent model through the use of a conceptual framework called Intelligent Collaborative Framework, addressing the gap between university computing graduates and the industry needs. This model is achieved via machine learning classifiers to recognize the issue and solve the problem between university computing graduates' and employers' expectations. In addition, the study identifies the required skills for computing graduate students to be employed in the industry. Several experiments were conducted using a dataset gathered from two computing departments and through a survey done among the graduates. The experiment results show that the ADA, SVM, and LR outperform the other classifiers. The model performance accuracy reached 89% for F1-Score. In addition, the best features (computing and training courses) were identified using the SelectKBest. The mutual information gain can assist in quickly obtaining jobs.

Keywords—machine learning; computing graduates; employability; industry skills

I. INTRODUCTION
The unemployment issue has arisen due to the mismatch between the labor market needs and the skills, competencies, and knowledge of university graduates [1-4]. The labor market requires high skills to be developed in graduates to cope with the fast technological advancement. For example, some governments in the gulf area enforce the employment of local people, and the companies fail that request due to the skills needed for vacant jobs [5, 6]. Many studies have highlighted the employability issue and have given attention to the gap between the insufficient supply of skilled graduates and the demand of the market in many countries such Sri Lanka [7]. USA [8], China [9-10], Tanzania [12], Europe [13-16], India [17], Ghana [18], Malaysia [6, 19-21], Russia [22], Vietnam [23], Nigeria [24], and Saudi Arabia [5, 11, 25-33]. The university curriculum needs to be updated in order to cover the technological advancements and required competencies [34, 35]. Authors in [17] highlighted the gap between the engineering graduates and the market needs in India by focusing on training the graduates in order to join the industry with market needed skills. In the same way, authors in [18] suggested conducting an awareness session about this issue.

Authors in [35] focused on closing the gap between software engineering graduates and the job industry’s needs, by proposing the adoption of better teaching methods (universities) on soft skills and better hiring efforts (employers). In [23], focus is given on the evaluation of university graduates from their managers based on graduate capabilities, skills and knowledge through surveys. Authors in [22] raise the attention on highly qualified scientific and technical professionals in the job market of robotics where the demands are not met. Authors in [21] highlighted the issue of employability for Malaysian graduates in the IR 4.0 industrial revolution. They attempted to solve this issue by proposing a framework called the Learning Factory (LF). In [36], a machine learning approach in Spain was conducted to test the predictions made for the changes in future job needs in order to prepare the students more adequately. Authors in [37] considered the gap between the job market of business analytics and the graduates. Authors in [38] proposed enhancing the graduates’ skills through authentic learning approaches in order to reduce the percentage of unemployment in Europe. Authors in [24] examined what the Labor Market (LM) actually demands from the Higher Education Institutions (HEIs) and how the demands of the LM can be met by the HEIs in Nigeria. Authors in [39] inspected the HEIs in Nigeria...
to get a better understanding about why there is a large percentage of unemployed graduates. Authors in [16] studied the relationship between learning and real practice in the industry finding a gap between them. Authors in [17] highlighted the impact of the curriculum and learning outcomes on graduates on the LM and investigated the relationship between higher education and industry. Moreover, authors in [18], suggest that the accounting curriculum in Saudi universities is required to be updated.

In Saudi Arabia, the Ministry of Higher Education has significantly improved the monitoring of the education process at a national level by introducing an organization called NCAA that monitors the progress of the academic programs in universities [13]. Practical courses make the graduates high skilled, as required by the industry. In addition, authors in [13] state that the graduates should have skills and knowledge and such elements have to add to the course learning outcomes. Such a way can be beneficial to the LM and authors in [19] introduce an appropriate framework. Techniques and presentation skills can be acquired through interpersonal and strategic competencies. Authors in [15, 20] demonstrate the importance of the English language courses that are fundamental in the job market. Authors in [40, 41] focused on unemployment issues and the mismatch of graduates with the market needs.

The main motivation for this study is the massive unemployment among university graduates [17, 18]. Limited attention has been given in enhancing and updating the curriculum of computer studies while no attention has been given in providing a comprehensive solution to solve the unemployment issue of graduates using artificial intelligence approaches. Based on our findings, a list of the desirable topics that should be reflected in the course plan of computer studies (computer science and information technology) was created. Besides, the experiments provide a full picture of the mismatch between the academic program characteristics, the employment potential, and the expectations of the employers. Furthermore, an intelligent collaborative framework bridging university computing graduates and the industry to provide valuable advice through a friendly visualization environment was created. This environment provides a list of desirable skills that are required in the LM to be in the computer science course plan. In such a way, an intelligent collaborative framework for bridging the gap between university computing graduates and the industry needs will reduce the unemployment rate of graduate students from computer science programs and give knowledge to the top management of that department of the required critical steps. The main contribution of this paper can be summarized in the following:

- The employability issue of computing students and the market needs are explored globally.
- A machine learning model is proposed to detect and identify if the graduate can get employed quickly or not. In addition, the impact of training courses on the computing graduate students is measured.
- A novel dataset for computing graduate students is introduced.

- Feature selection methods are applied on the proposed model and the employability dataset.
- The results of commonly used machine learning classifiers are compared.

II. MATERIALS AND METHODS

This section explains the conceptual framework and the phases, data collection, data cleaning, filtering, machine learning model, and model evaluation, that are were used to carry out this study (Figure 1).

A. Data Collection

This is the initial phase of the proposed model. In this phase, the data collected are divided into 3 stages. In the first stage, five experts examined the core course of computing based on the international standards, benchmarks from international universities, and the curriculum of computer science of four universities. The core courses were identified and listed. Based on these courses, in the second stage, computing graduate marks were collected from graduates from two universities, 2-5 years after their graduation. In the third stage, a survey was sent to these graduates and the results were analyzed. The total number of surveyed graduates was 486 while the number of received surveys was only 277 as shown in Figure 2. This process is performed annually as a procedure of academic quality and development.
Figure 3 illustrates the number of graduates that took and did not take training courses. Figure 4 shows the statistics on the total number of graduates that took training courses in Machine Learning (ML), Data Science (DS), Natural Language Processing (NLP), and Artificial Intelligence (AI). The majority of courses that were taken by the graduates were ML and AI.

B. Data Cleaning and Filtering

As an input from the previous phases, the data have been cleaned. In addition, the data of graduate students were merged with the graduate survey. The dataset was constructed at this stage with 29 features, from which 25 are computing courses and 4 are training courses. The average marks of the computing courses of the graduates are shown in Figure 5.
C. Baseline Machine Learning Classifiers

The utilized ML classifiers were identified in this stage, based on their popularity. These classifiers are k-Nearest Neighbors (kNN), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Naive Bayes (NB), Multinomial Naive Bayes (MNB), Stochastic Gradient Descent (SGD), AdaBoost (ADA), and Gradient Boosting (GB) [42, 43].

D. Feature Selection Methods

In order to reduce the number of features and to identify the most important among them that effect employability, feature selection took place. The most common techniques used are SelectKBest and mutual information gain. In SelectKBest, selection is conducted on the highest correlated features to the independent variable (Class/Y) using the chi2 score function. The mutual information is a statistical method used to calculate the dependence between two features and then sort them in order to select the most correlated features.

E. Model Evaluation

This section describes the most common methods used to measure the model performance. These measures are Precision, Recall, F1-Score, and Accuracy [44] and are used in this study. The relevant equations are (1)-(4). In addition, the Area Under the ROC Curve (AUC) and confusion matrices were used.

\[
\text{Precision} = \frac{TP}{TP+FP} \tag{1}
\]

\[
\text{Recall} = \frac{TP}{TP+FN} \tag{2}
\]

\[
\text{F1 – Score} = \frac{2 \cdot (\text{Precision} \cdot \text{Recall})}{\text{Precision} + \text{Recall}} \tag{3}
\]

\[
\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}
\]

where TP stands for True Positive, TN for True Negative, FP for False Positive, and FN for False Negative.

III. RESULTS AND DISCUSSION

This section presents the results of the carried-out experiments based on the common used ML classifiers to identify the employability of computing course graduates and if the training courses influence hiring after graduation. These experiments were based on a set of settings and hyper-parameters. The ML classifiers used were ADA, DT, GB, kNN, LR, NB, NB, SGD, SVM, RF, and XGB. The experimental results are shown in Table II. All the experiments were carried out using Keras on Tensor flow architecture and Python programing language. The SKlearn library was utilized to train the proposed model using the aforementioned ML classifiers. Figure 6 demonstrates the results of the experiments and shows the model performance in terms of Accuracy. The experimental results show that LR and SVM scored the highest recorded accuracy of 87% and 86%, respectively, while kNN and SGD scored the worst with 48% and 54%, respectively. Also, SVM, LR, and GB scored the highest average Precision and Recall values, ranging between 85% to 88%, while kNN had the worst score of 47% as shown in Figures 7 and 8. The AUC is shown in Figures 9 and 10.

<table>
<thead>
<tr>
<th>ML Classifiers</th>
<th>Classes</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADA</td>
<td>0</td>
<td>0.82353</td>
<td>0.76231</td>
<td>0.821918</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.816327</td>
<td>0.90909</td>
<td>0.86215</td>
</tr>
<tr>
<td>DT</td>
<td>0</td>
<td>0.789474</td>
<td>0.76231</td>
<td>0.779221</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.8</td>
<td>0.81818</td>
<td>0.808989</td>
</tr>
<tr>
<td>GB</td>
<td>0</td>
<td>0.909091</td>
<td>0.76231</td>
<td>0.833333</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.82</td>
<td>0.93182</td>
<td>0.87234</td>
</tr>
<tr>
<td>KNN</td>
<td>0</td>
<td>0.441176</td>
<td>0.38461</td>
<td>0.410959</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.510204</td>
<td>0.58618</td>
<td>0.537634</td>
</tr>
<tr>
<td>LR</td>
<td>0</td>
<td>0.891892</td>
<td>0.84615</td>
<td>0.868421</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.869565</td>
<td>0.90909</td>
<td>0.888889</td>
</tr>
<tr>
<td>MNB</td>
<td>0</td>
<td>0.909091</td>
<td>0.76231</td>
<td>0.833333</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.82</td>
<td>0.93182</td>
<td>0.87234</td>
</tr>
<tr>
<td>NB</td>
<td>0</td>
<td>0.9</td>
<td>0.69208</td>
<td>0.782609</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.77585</td>
<td>0.93182</td>
<td>0.845361</td>
</tr>
<tr>
<td>SGD</td>
<td>0</td>
<td>1</td>
<td>0.02561</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.536585</td>
<td>1</td>
<td>0.69841</td>
</tr>
<tr>
<td>XGB</td>
<td>0</td>
<td>0.861111</td>
<td>0.794872</td>
<td>0.826667</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.820787</td>
<td>0.886364</td>
<td>0.857143</td>
</tr>
<tr>
<td>RF</td>
<td>0</td>
<td>0.903226</td>
<td>0.717949</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.788462</td>
<td>0.93182</td>
<td>0.854167</td>
</tr>
<tr>
<td>SVM</td>
<td>0</td>
<td>0.886889</td>
<td>0.820513</td>
<td>0.853333</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.851064</td>
<td>0.90909</td>
<td>0.879121</td>
</tr>
</tbody>
</table>
IV. CONCLUSION

This paper proposed machine learning models to identify whether computing graduates have been employed after graduation or not. The models were trained on a dataset that was introduced and validated by experts in the area of computing. This problem is considered as a binary classification with two classes, i.e., employed and not employed. The machine learning classifiers were trained and assessed with Precision, Recall, F1-Score and Area Under the ROC Curve. In addition, feature selection techniques were utilized in order to determine the most important courses (features) that can affect the employability of university graduates in the areas of computing. The highest model performance was achieved with kNN, LR, and VSM and reached 89%.

The experimental results show that the computing curriculum must be extended with more practical courses and topics about emerging technologies such as dataset science, machine learning, and natural language processing. In the future, the dataset could be extended and more records of graduates could be used. Also, a visual representation of the recommendations could be designed.


