Utilizing Machine Learning Algorithms for Recruitment Predictions of IT Graduates in the Saudi Labor Market

Munirah Alghamlas[1] and Reham Alabduljabbar[2]
[1]Information Technology Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia, 4
[2]Information Technology Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia,

Abstract
One of the goals of the Saudi Arabia 2030 vision is to ensure full employment of its citizens. Recruitment of graduates depends on the quality of skills that they may have gained during their study. Hence, the quality of education and ensuring that graduates have sufficient knowledge about the in-demand skills of the market are necessary. However, IT graduates are usually not aware of whether they are suitable for recruitment or not. This study builds a prediction model that can be deployed on the web, where users can input variables to generate predictions. Furthermore, it provides data-driven recommendations of the in-demand skills in the Saudi IT labor market to overcome the unemployment problem. Data were collected from two online job portals: LinkedIn and Bayt.com. Three machine learning algorithms, namely, Support Vector Machine, k-Nearest Neighbor, and Naïve Bayes were used to build the model. Furthermore, descriptive and data analysis methods were employed herein to evaluate the existing gap. Results showed that there existed a gap between labor market employers' expectations of Saudi workers and the skills that the workers were equipped with from their educational institutions. Planned collaboration between industry and education providers is required to narrow down this gap.

Keywords:
Classification, data mining, machine learning, labor market, prediction, Saudi Arabia, skills gap, recruitment

1. Introduction

Information Technology (IT) has recently become indispensable for any country's economic growth. IT knowledge and skills are critical elements for nations to prosper and compete. Therefore, the quality of education and ensuring that graduates possess sufficient knowledge and skills to meet the demands of the market are necessary. The gap between the educational output and the actual demand of the labor market could hinder Saudi Arabia in becoming a competitive nation, in terms of IT solutions, that provides IT services to regional and global markets. However, Saudi Arabia is one of the top spenders in the information and communications technology. In 2017, Saudi's spending in IT reached to $14.2 billion (MCIT, 2014). Such a massive spending leads to the creation of a large number of IT jobs within the IT departments of organizations and in local IT companies. Conversely, one of the real challenges of the 21st century in Saudi Arabia is the high rate of unemployment. In 2020, the unemployment rate is increasing according to the statistics from the Saudi labor market. One of the causes behind this high rate is the gap between the education outputs and the actual requirements of the Saudi labor market ("Labor market reports," 2020).

Nowadays, with the internet evolution, employers and job seekers are exploring online portals to find employees and jobs, respectively. As information about the requirements of the labor market in Saudi Arabia is available on online job portals and the current education outputs can be extracted from the users, we can match the demands of the Saudi labor market to solve the problem of the high rate of unemployment by knowing the disequilibrium among current education outputs, particularly in IT jobs.

There are two main contributions herein. The first contribution is the development of a prediction model that can be deployed on the web, where users can input variables to get predictions on their suitability for recruitment in the Saudi IT labor market. The model can generate predictions directly from the source, that is, by using data extracted from two online job portals, namely, LinkedIn and Bayt.com, as posted by employers in their job announcements. The second contribution is providing data-driven recommendations of the in-demand skills in the IT labor market in Saudi Arabia to overcome the unemployment problem. This study seeks to resolve the following research questions:

- What are the most in-demand skills required for recruitment in the IT Saudi labor market from the employer's point of view?
- How can we predict the suitability of an IT graduate for recruitment in the IT Saudi labor market?
- How can we reduce the perceived skills gap and enhance the employability of IT graduates in the IT Saudi labor market?

Manuscript received March 5, 2024
Manuscript revised March 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.3.14

2. Background and Related Works

This section comprises three subsections: (A) research on the skills gap and unemployment in Saudi Arabia, (B) available online job portals, and (C) research on prediction models and their applications for employment.

2.1 **Skills Gap and Unemployment in Saudi Arabia**

Skills gap is known as the difference between the skills required for a job versus those skills that a prospective worker possesses (ACT, 2011). One of the main obstacles is that many of the graduates are not equipped with industry-relevant skills, particularly in industries related to science and technology. In Saudi Arabia, recent research shows that Saudi's ICT sector exhibits skills gap in two primary areas: technology-based skills and soft skills. It is important to note that many of the technology-based skills are taught at college or university level (Mishrif & Alabduljabbar, 2018). Also, in Saudi Arabia, there is a lack of coordination between institutions of higher education and the world of work (Ibeaheem, Ragmoun, & Elawady, 2017). Moreover, recent evidence from 2018 shows that this gap still exists. Hence, the IT industry must work with the Ministry of Education in Saudi Arabia to equip students with the required skills. In general, most employers believe that graduates from the universities are not unequipped with the required soft skills (Kennedy, 2019).

In addition, as mentioned earlier, one of the real challenges of the 21st century in Saudi Arabia is the high unemployment rate. The latest General Authority for Statistics (GaStat) labor market release shows that the unemployment rate jumped to 15.4% in Q2 2020 from 11.8% in Q1 2020, as shown in Fig. 1("Labor market reports," 2020).The statistics shows that Saudi Arabia still suffers from a high unemployment rate.

Saudi Unemployment Rate
(percent)

	Q2 2020	Q1 2020
Male	8.1	5.6
Female	31.4	28.2
Youth (20-24)	35.4	31.3
Total	**15.4**	**11.8**

FIGURE 1. Unemployment rate in Saudi Arabia Q1 and Q2, 2020

Skills gap is a worldwide concern; hence, researchers have studied this issue from different perspectives in the literature. The main objective of these studies was to determine the most important skills or jobs in the IT labor market. A common aspect among all of these studies is that they collect and analyze data to evidence that the skills gap exists.

In Saudi Arabia, one study that provides a comprehensive analysis of the ICT skills required for employment in local ICT organizations and it identifies the gap between this requirement and the ICT skills cultivated by the educational institutions of Saudi Arabia. An interview was conducted with approximately 200 ICT workers who were selected randomly from the market. Results show that 80% of the employers require Java programming skills, 70% require Visual Basic, and 60% of the employers require programming skills in C++, Java Script, and Visual Basic Script. Also, the ICT worker survey indicated that the highest skill required in their jobs is English technical writing skills (Alsafadi & Abunafesa, 2012). Moreover, a report curated by the ICT workforce in Saudi Arabia shows that there is a need to focus on the developmental skills related to cloud infrastructure, software development, security, mobility, ICT strategy, IT administration, helpdesk, and support (CITC, 2015).

In United States, four studies were conducted to understand the required IT skills using online job advertisements. In the first study, the researchers extracted almost a quarter million unique IT job descriptions from various job search engines and distilled each to its required skill sets. Using their own web content data mining application, they revealed 20 clusters of similar skill sets that map to specific job definitions. Their study aims to allow software engineering professionals to improve their skills to match with those in demand from real computing jobs across the US that the study tested (Aken, Litecky, Ahmad, & Nelson, 2010). The second study aims to identify the skills that entry-level IT consultants require. Focus groups from different companies and open-ended surveys for senior IT student are used to gather more in-depth data. Results show that both written and oral communication skills were essential and that senior consultants looked for well-rounded entry-level individuals (Lending & Dillon, 2013).

The third study (Gallivan, Truex, & Kvasny, 2002) examines the trends associated with the required job skills for IT professionals. Through an empirical study of a dataset comprising job advertisements for IT professionals over the past 13 years, 1691 job advertisements were captured and analyzed. This study then evaluated whether the observed trends support earlier predictions offered by researchers who sought to anticipate future job and skill demands. They also looked for evidence of whether the gap between employers' demands and the skills provided by academic programs still exists. Results did not obtain any evidence of a recruitment gap. Also, it was suggested that

the job advertisements focus on "hard skills" (Gallivan et al., 2002).

In the last study, a research was conducted to examine the IT job skills across three genres of texts: scholarly articles, practitioner literature, and online job advertisements (241 online job advertisements listed on Monster.com from April 2008–June 2008). Three datasets were analyzed in two stages. In the first stage, three coding schemes were developed based on the articles and job advertisements. The codes were classified into three standard categories: humanistic skills, business skills, and technical skills. In the second stage, the researchers summarized the data across the three data sources to develop a comprehensive set of code (Huang, Kvasny, Joshi, Trauth, & Mahar, 2009). Results show that the online advertisements list a strong mix of skills in these three categories, whereas the literature concentrates on technical skills. Table I presents a summary of all of these studies.

2.2 Online Job Portal

Online job portals represent a central component of the modern employability industry. Job seekers browse through job announcements every day. On the other hand, companies can announce their openings and search through requests and CVs of potential employees. Two popular online job portals in the middle east are used for data collection herein: LinkedIn and Bayt.com.

1) LINKEDIN

LinkedIn is a social networking site that is specifically designed for the business community with 675 million users ("Linkedin," 2020). The goal of the site is to allow registered members to establish and document networks of people they know and trust professionally. On LinkedIn, people can also seek jobs or add a job advertisement. LinkedIn started from the west in the United States in 2003; however, at the same time, it is growing rapidly all over word, including the Middle East.

2) BAYT.COM

East Bayt.com is the leading online recruitment website in the Middle East ("Bayt.com," 2020) with more than 30 million registered job seekers and over 10 million visits each month. Imagine the amount of the interesting and powerful information that can be acquired from online job portals like Bayt. Thousands of new jobs are listed daily on Bayt platform from the region's top employers. Jobs are classified on the basis of qualifications, salary, and requirements. Established in Dubai in 2000, Bayt has 12 regional offices that are spread throughout the Middle East, in Kuwait, Amman, Cairo, Doha, Dubai, Jeddah, Riyadh, and Abu Dhabi.

2.3 Prediction Model Using Machine-Learning Algorithms in Employment

Machine learning is a branch of artificial intelligence based on the notion that systems can learn from data, identify patterns, and make decisions with minimal human intervention. Any system can be trained to learn from data. Machine learning algorithms characteristically fall into one of the two learning types: supervised and unsupervised learning. Supervised learning refers to working with a set of labeled training data. For every example in the training data, you have an input object and an output object (known as the input and output data). Unsupervised learning is where you let the algorithm determine a hidden pattern in a load of data. With unsupervised learning, there is no definite right or wrong answer. It is just a case of running the machine learning algorithm and observing what patterns and outcomes occur (Bell, 2014).

Normally, supervised learning techniques are used for classification models. Several algorithms were proposed for the supervised classifications of texts. Among these algorithms, Support Vector Machine (SVM), Naïve Bayes (NB), and k-Nearest Neighbor (k-NN) were shown to be the most appropriate in the existing literature (Khan, Baharudin, Lee, & Khan, 2010), which are investigated herein to build the predictive model.

Previous studies have reported the use of prediction models in employment. Some models are used to match the right talent to the right job (Zhu et al., 2018), whereas others predict employee turnover (Bhulai, 2016). The Person-Job Fit model (Zhu et al., 2018) matches the right talent to the right job by identifying talent competencies that are required for the job through the use of a novel end-to-end data-driven model based on convolutional neural network (CNN), namely Person-Job Fit Neural Network (PJFNN). PJFNN not only estimates whether a candidate fits a job, but it can also identify which specific requirements in the job posting are met the candidate. Also, the model is evaluated based on a large-scale real-world dataset collected from a high-tech company in China.

Another predictive model was built for predicting a career success in engineering among women and African American men (Charitable, 2011). Data were collected

through a survey size of 400 successful engineers and 400 engineering dropouts (women and African American men) and analyzed in a systematic manner to gain a deeper understanding of the underrepresentation of women and African American men in engineering. Results show that the model was concordant by 99.4% with the real world. Also, another research (Paparrizos, Cambazoglu, & Gionis, 2011) studied the problem of recommending jobs to people who are seeking a new job. They developed a supervised learning model that predicts the next job transition of a person and recommends jobs to people. The prediction model was trained using a large number of job transitions obtained from the web and by predicting the next institution for an individual. If the accuracy of such prediction is sufficiently high, the model will recommend institutions to employees who are seeking jobs.

Another valuable study (Bhulai, 2016) used three different predictive models to analyze the factors that influence the prediction of employee turnover in an organization using data mining techniques. The aim of this study is to provide recommendations that can enhance the efficiency and effectiveness of human resource planning processes that are used to focus on the employee-turnover problem. This study is conducted on real data that was provided by Focus Orange Technology in Amsterdam. The data comprises 29 attributes and 10,616 instances, which was used for training and testing purpose. The selected models are logistic regression, artificial neural networks, and random forest. Results indicate that the random forest works best on these datasets. A similar study (Punnoose & Ajit, 2016) aimed to compare the extreme gradient boosting (XGBoost) technique against six supervised models using the data from HR Information Systems (HRIS) to predict employee's turnover. Results show that (XGBoost) provided significantly higher accuracy in predicting the employee turnover. Table II summarizes the mentioned models.

In short, the review of the literature emphasizes shortage of studies in Saudi Arabia. Accordingly, this study aims at providing a link between IT graduates and the labor market. The work presented herein has the following key features that distinguish it from other works in this area: collecting and analyzing data by leveraging two online job portals Bayt.com and LinkedIn using the collected data to identify the actual requirements of the IT skills in recruitment in Saudi labor market; building a prediction model to predict the suitability of IT students' skills for the recruitment in Saudi labor market; providing a set of recommendations that may contribute to the reduction of the unemployment rate in general and to meet the requirements of the IT Saudi labor market.

TABLE II
SUMMARY OF MODELS USING MACHINE LEARNING TECHNIQUES IN EMPLOYMENT

Model name	Methods/ Models used	Data Type	Evaluation	Reference
PJFNN	Logistic Regression (LR), Decision Tree (DT), Naive Bayes (NB), Support Vector Machine (SVM), Ad boost (Ada), Random Forests (RF), Gradient Boosting Decision Tree (GBDT), Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA).	Textual data	Yes, 10% of the dataset as test data	(Zhu et al., 2018)
Predicting the employee turnover	Three types of model (logistic regression (LR), artificial neural networks (ANN), and random forest (RF)).	Numeric and textual data	Yes, k-fold cross validation	(Bhulai, 2016)
Career Success in Engineering	Correlation analysis used to analyze the data and logistic regression (LR) model was developed to determine the probability of staying in engineering or leaving the field.	Numeric data	Using chi-square	(Charitable, 2011)

Job Recommendation model	Several machine-learning algorithms; however, they present only the algorithm for the decision table/naive Bayes hybrid classifier (DTNB), which achieves the highest accuracy.	Textual data	Yes, Part of the dataset as test data	(Paparrizos et al., 2011)
Employee Turnover in Organizations	XGBoost, Logistic Regression (LR), Naïve Bayesian (NB), Random Forest (RF), SVM (RBF kernel), and KNN.	Numeric	Yes, k-fold cross validation	(Punnoose & Ajit, 2016)

3. Methodology

This section is divided into two: the problem is formulated in section A, and a detailed explanation of the approach is presented in section B.

3.1 Rational and Overview

Consider that a typical job seeker k has a set of job seeker skills S, where S ={s1, s2,…, sn}, and the Saudi labor market has a set of wanted skills WS, where WS = {ws1,ws2,…wsn}. The suitability ST of k in the Saudi labor market can be calculated herein as

$$ST = (s / ws) \times 100$$

This work seeks to address the following research questions:

- What are the most in-demand skills required for recruitment in the IT Saudi labor market from the employer's point of view?
- How can we predict the suitability of an IT graduate for recruitment in the IT Saudi labor market?
- How can we reduce the perceived skills gap and enhance the employability of IT graduates in the IT Saudi labor market?

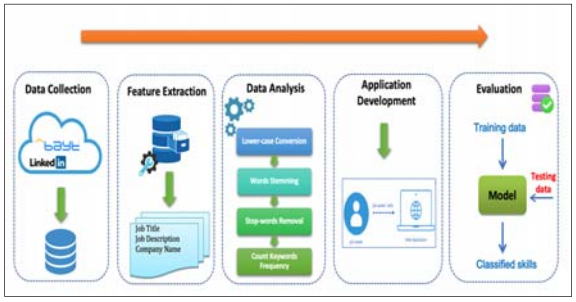
The first research question has been addressed by collecting and analyzing data from online job portals to better understand the in-demand skills and jobs. To resolve the second research question, a machine-learning algorithm was utilized to build the suitability prediction model and a web-based application was developed to interact with the model based on the model. Finally, to evaluate the existing gap in education, senior and graduate IT students were surveyed to determine whether they possess the required skills as predicted from our analysis on job advertisements on LinkedIn and Bayt. The overall solution can be seen in Fig. 2. The data was collected from LinkedIn and Bayt.com. The data set was then preprocessed by running some filters for cleaning the data set. The data set was then labeled, and the skills were classified into two categories: Wanted Skills and Unwanted Skills. Next, a machine-learning algorithm was used to build the prediction model. The model was trained to classify new skills into Wanted Skills and Unwanted Skills. During this phase, the model was tested and evaluated. Furthermore, a web app was created so that job seekers can interact with the trained model, i.e., they could input their skills to predict their suitability for the Saudi labor market. Finally, to evaluate the existing gap in education, senior and graduate IT students were surveyed to determine whether they possess the required skills as predicted from our analysis on job advertisements on LinkedIn and Bayt. A detailed explanation of each phase is provided in the following subsections.

FIGURE 2. Methodology

A. Dataset Preparation

1) DATA COLLECTION

Data were collected from two online job portals,



LinkedIn and Bayt.com. To post jobs on job portals, requesters have to specify the job features, e.g., title, description, and required skills. Among these features, job location, job function, job title, and job description were the most relevant as per the context of our job. Data were collected over a three-month period (October 2018–December 2018) through web scraping using the OCTOPARSE tool, as shown in Fig. 3 ("Octoparse," 2020).

OCTOPARSE is an automatic powerful tool used to extract data from web pages.

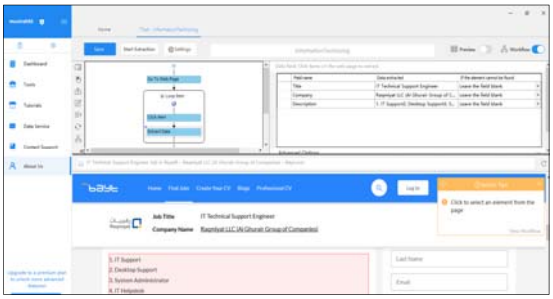


FIGURE 3. Data extraction from Bayt.com

Overall, the total number of job posts extracted was 675 job requests with each job request having its own features, e.g., title, description, and required skills. Fig. 4 shows a part of the data set. Subsequently, a content analysis was performed on these job requests as a first step to determine the most requested skills within these job requests. Content analysis is a widely used qualitative research technique (Krippendorff, 2018). Each job request was processed by the tool ("Octoparse," 2020) as a separate case. The resulted data set comprised 170 skills.

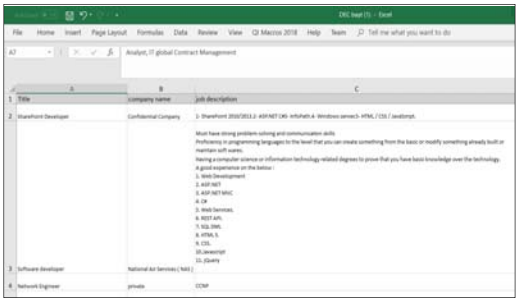


FIGURE 4. Part of the dataset

2) DATA CLEANING

To meaningfully label the dataset, preprocessing of the dataset was accomplished by applying some filters using the OpenRefine tool (OpenRefine, 2020) (see Fig. 5) in the following order:

1. Remove unwanted characters (nonalphabetic characters) using the "expression value" option in the OpenRefine.
2. Delete duplication.
3. Stop-words removal filter: Stop words are words that are filtered out before the processing of textual data. Some of the common stop words are ''the, is, at, but be, been, and, as, out, ever, own, he, she, and an".
4. Filtering/Faceting Data: It is a method to filter data into subsets for ease of use that can be done for text, number, and dates. (Facet→Text Filter).
5. Lowercase filter: Keywords are converted to lowercase letters to simplify the comparison. After the cleaning process, the dataset was pared down to 120 skills.

FIGURE 5. Data cleaning using OpenRefine Tool

3) DATA LABELING

To classify the data set into two classes, "wanted" and "unwanted" skills, the average of the occurrences of skills was used as a determination point. Any skill with an occurrences number above the average was considered as a "wanted" skill, whereas any skill with an occurrences number below the average was considered as an "unwanted" skill. After experimenting, the average was calculated using the median formula because it is the most suitable measure of average for data classified on an ordinal scale. Also, the median formula is a good measure of the average value when the data include exceptionally high or low values (SSDS, 2010). The skills were first arranged in ascending order, and the median is calculated using the following formula:

$$\textbf{Median} = \{(n + 1) / 2\}^{th} \text{ value}$$

where n = number of skills,

As there is an even number of skills in the dataset (120), there is no longer a distinct middle value. The median is the 60.5^{th} value in the data set, which suggests that it lies between the 60^{th} and 61^{st} value. By averaging the 60^{th} and 61^{st} value of the dataset, the calculated median was 27 and it becomes the determination point. Thus, all skills that have occurrences of more than the determination point (27) are wanted skills, whereas the rest are classified as unwanted skills. Table III shows part of the data set. As observed, "Statistical skills" was classified as "unwanted" because its occurrence number is below 27, whereas "Analysis Skills" was classified as "wanted" because its occurrence number

is above 27. The distribution of the wanted and unwanted skills within the data set in Fig. 6 shows that 50% of the skills are wanted and 50% are unwanted skills.

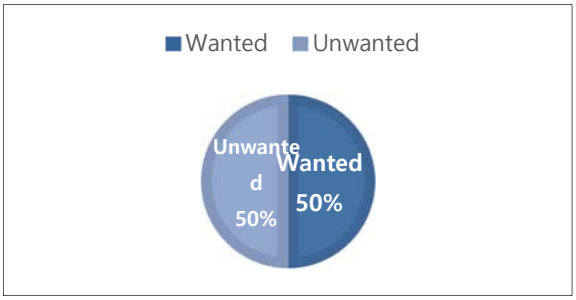


FIGURE 6. Dataset distribution

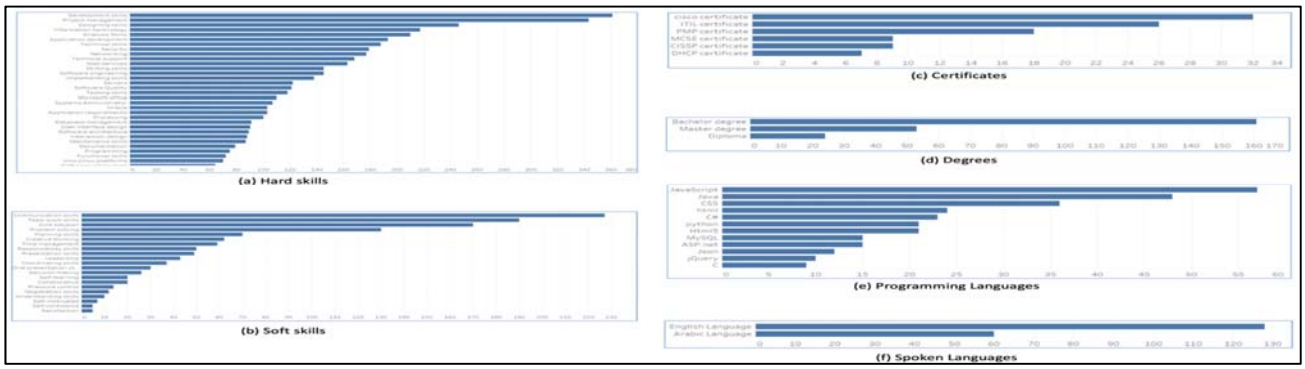


FIGURE 7. In-demand skills in each skill category

4. Implementation and Results

4.1 Experiment 1. Answering the First Research Question: What are the Most In-Demand Skills Required for Recruitment in the IT Saudi Labor Market from the Employer's Point of View?

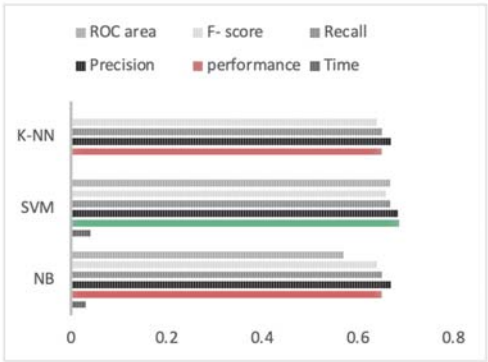
To better understand in-demand skills for recruitment in the IT Saudi labor market, the related skills are grouped into categories. Skills categorization was created by examining the Association for Computing Machinery (ACM) IT Curriculum (ACM, 2017) and other empirical study (El-Gabaly & Majidi, 2003). Accordingly, skills were categorized into hard skills, soft skills, programming languages, study degree, certificates, and languages. Subsequently, a list of 76 hard skills, 21 soft skills, 13 programming languages, three study degrees, six certificates, and two spoken languages' skills were identified. Fig. 7 provides a summary of the in-demand skills in each skill category.

Among the hard skills shown in Fig. 7(a), the development and project management skills are the most required skills, whereas among the soft skills listed in Fig. 7(b), communication and teamwork skills are the most required in the Saudi labor market from October to December 2018. Fig. 7(c) shows that the most required certifications in IT are CISCO and ITIL, whereas Fig. 7(d) shows that the most required educational degree in IT jobs is a bachelor's degree. Moreover, Fig. 7(e) shows that the most in-demand programming languages are JavaScript and Java, whereas Fig. 7(f) shows that most IT jobs require fluency in English.

4.2 Experiment 2. Answering the Second Research Question: How Can We Predict the Suitability of IT Graduates for Recruitment in the IT Saudi Labor Market?

Normally, supervised learning techniques are used in classification models. Among the several algorithms proposed for the supervised classifications of texts, Support Vector Machine (SVM), Naïve Bayes (NB), and k-Nearest Neighbor (k-NN) were used in this study because these algorithms were shown to be the most appropriate in the existing literature (Khan et al., 2010).

The resulting dataset was used as a training sample for the classifier to automatically detect the class of unlabeled skills. With this small data size, it is recommended to use n-fold cross-validation and percentage split technique (Brownlee, 2018). N-fold cross-validation implies that the data set is split into n (10 for this study) partitions. Then, the first n−1 parts are used for training, and the nth part is used



for testing. This process is repeated allowing each of the n parts of the data set an opportunity to be the test set. For the percentage split technique, 70% of the data is used to train the model and 30% is used for testing.

The classification was conducted using the Waikato Environment for Knowledge Analysis (Weka) toolkit. It is a suite of machine-learning software written in Java, developed at the University of Waikato, New Zealand. It is a free software licensed under the GNU General Public License (Waikato, 2016). To be able to load the data set into WEKA, the data set format was converted from CSV to ARFF format. As the data type is text, "Filtered Classifier" algorithm and "StringToWordVector" filter must be used to process each string into a vector of word frequencies for further analysis with different data mining techniques before selecting the classifier technique to convert the class to 0 and 1 (Hamoud & Atwell, 2016). Before deciding on which classification algorithm to use, an experiment was conducted with other classifiers, namely, NB and k-NN. Fig. 8 shows a comparison between the three classifiers using the10-fold cross-validation.

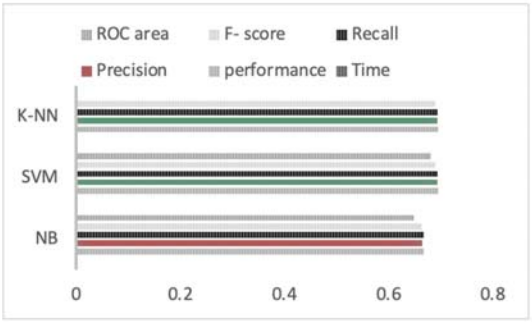


FIGURE 8. Comparison of classifiers using the 10-fold cross-validation

Comparison results show that SVM achieved a good performance in comparison with the k-NN and NB classifiers using the 10-fold cross validation. However, determining the accuracy alone is insufficient, as sometimes, having high accuracy values does not necessarily imply that the algorithm has an excellent performance. Hence, it is necessary to look at other aspects to decide which algorithm is the best. Percentage split (70%) validation technique was used to ensure that the model performs well with different test data in the future. Fig. 9 shows a comparison between the three classifiers using percentage split validation.

FIGURE 9. Comparison of classifiers using percentage split

Results from both comparisons (Fig. 8 and Fig. 9) showed that SVM was the best classifier technique. With percentage split option, K-NN and SVM yielded the same accuracy, which was 69.44%; however, ROC area in SVM was better than that in K-NN. Hence, SVM was the best classifier with the highest accuracy.

To enhance the model performance, domain experts were consulted who came up with four options that can be opted in the study. The options include increasing the dataset size, performing feature selection, changing classifier parameter, and using ensemble methods (Bagging and Adaboots). Increasing the size of the data takes a long time, which is not desired herein. The feature selection option has been eliminated as well because it is not recommended to apply attribute selection to all of the data and then run an evaluation on the dimensionally reduced data. Doing this will give overly optimistic error rates because the attribute selection process has used the data from the test folds.

Conversely, ensemble methods do not work perfectly with all the three classifiers. This leaves one available option, which is changing the classifiers parameter. In the SVM classifier, after changing the kernel function from PolyKerne to PUK, the accuracy increased to 72.22%, whereas no increase in the accuracy was observed in K-NN and Naïve base after changing the parameter. Fig. 10 depicts the incorrectly classified instances using SVM after enhancing the accuracy.

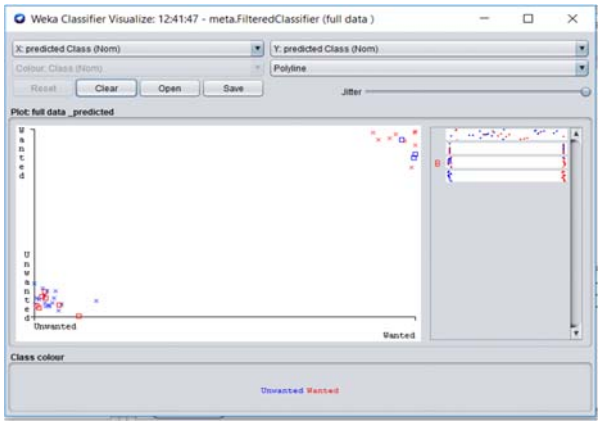


FIGURE 10. Incorrectly classified instances using SVM after increase the accuracy

1) DEPLOYING THE MODEL TO THE WEB

A web application was created so that the users can interact with the prediction model, i.e., the users can input their skills to generate predictions about their suitability of recruitment in the IT Saudi Labor market. A Java-based desktop application was implemented to facilitate the web interaction between the users and the proposed model. The desktop application was developed using the Eclipse

IDE (Eclipse, 2016) via the Weka library. The prediction model can be reloaded later on the desktop application and used exactly as if it had been trained (see Fig. 11). In other words, the desktop application can be utilized for classifying new dataset of skills in ARFF format using the study's prediction model. The results are automatically saved as a text file.

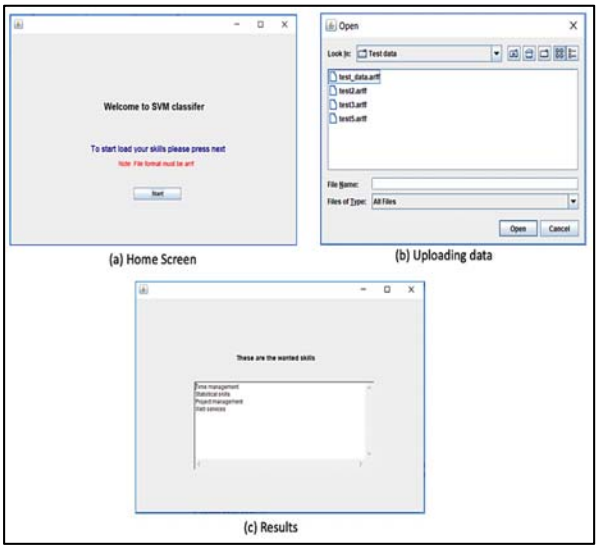


FIGURE 11. Screenshots of the desktop application

The resultant text file can now be used in the web application. Fig. 12 shows the web application interface and its different components. The interface was designed to be user-friendly, simple, and consistent. As shown in Fig. 12, the required skills are presented to the users. If the user fills in the form on the page and clicks on a "next" button, the application extracts the input, runs it through the model, and will finally render result.html with the results in place. The user can then view two items in the results page:

1. User's suitability for recruitment in the IT Saudi labor market.
2. Recommendations to help the user addresses her skills shortage.

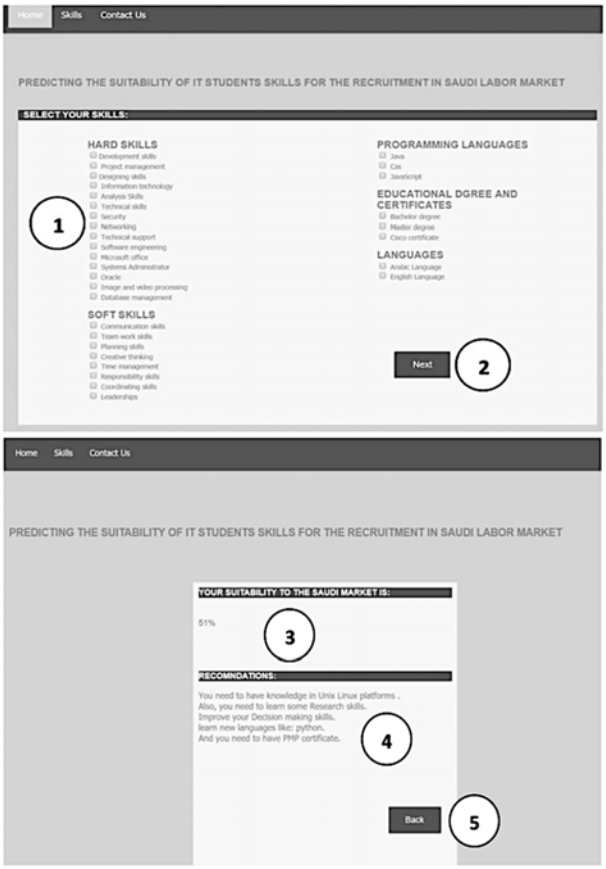


FIGURE 12. Web application interface: (1) User's skills selecting page; (2) Next button to the results page; (3) User's suitability to recruitment in the IT Saudi labor market; (4) Recommendations to help the user develop her skills; and (5) Back button to the skills page.

4.3 Experiment 3. Answering the Third Research Question: How Can We Reduce the Perceived Skills Gap and Enhance the Employability of IT Graduates in the IT Saudi Labor Market?

Descriptive and data analysis methods were employed in this study to evaluate the existing gap in education. Senior and graduate IT students in Saudi Arabia were surveyed to determine whether they possess the required skills as predicted from our market analysis of the online job portals. To ensure that the questions were clear, complete, and unambiguous, experts were asked to participate in a pilot study to validate the content and style of the survey. Feedback was gathered, leading to an improved version of the survey that was then distributed to the participants. The final survey comprised 44 skills and was administered via a commercial online survey tool. The

survey instrument comprised five main sections: demographic information, hard skills, soft skills, programming languages, and specialized technical certificates. Participants were asked to rank themselves in terms of importance on a scale of 1 (no level of competence) to 5 (high level of competence).

1) WEIGHTED AVERAGE

Incorporating Rating Scale questions, weights can be assigned to each answer option. This will enable to calculate the weighted average for each answer option. On the scale, the answer with a high score or high average is the most preferred one. The weighted average is calculated as follows,

$$\frac{x_1w_1 + x_2w_2 + x_3w_3 \ldots x_nw_n}{Total\ reponse\ count}$$

With Likert and Likert-like survey questions incorporating numbers as options, it is easier for participants to provide their answers and for researchers to analyze results. In relation to our data, any measured skill scored a weighted average above than three (the midpoint of the scale 1–5) can be considered as gained by the participant, whereas a value below three would indicate lacking this skill (Boone & Boone, 2012).

Responses were received from more than 100 participants, 72% of which were females and 28% were males. 79% of the respondents surveyed were senior IT students and 21% were IT graduates. The participants belong to different universities in Saudi Arabia, namely, King Saud University, Imam Mohammad University, Prince Nora University, and Prince Sultan University. Tables IV, V, and VI present the results collected from the students, which show their self-assessment of hard skills, soft skills, and programming languages, respectively, that they are either holding or lacking. Skills possessed by a participant with a weighted average greater than three are shown in bold.

Results show that the surveyed students claimed that they are lacking some of the required hard skills such as image and video processing and Unix Linux platforms, whereas they claim mastering the development, research, and analysis skills. Furthermore, all surveyed students claim that they have all of the soft skills. Regarding the programming languages, the students claim that they lack knowledge in crucial programming languages such as C# and Python. Fig. 13 shows that only 21% of the respondents hold IT certifications.

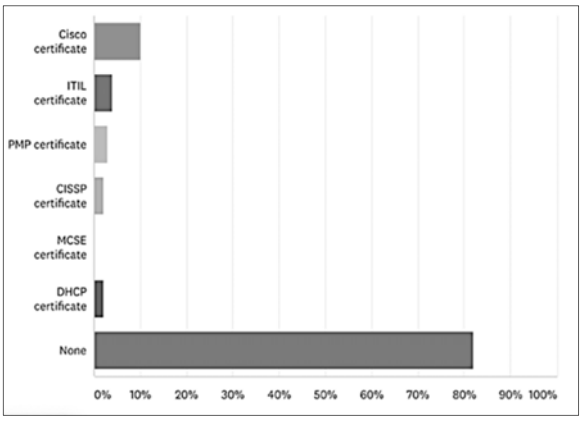


FIGURE 13. Certifications possessed by participants

5. Discussion

The first set of findings herein indicates that the top in-demand hard skills (Fig. 7(a)) by the Saudi IT market are development skills, project management, and analysis which are the same as the top-possessed skills by the senior IT and graduates survey (Table IV). In addition, the survey of IT graduates shows that mastering or familiarity with Microsoft Office is on the top of the acquired skills by an IT graduate; however, it was not at the top of the in-demand skills required by the market. This might be because as an IT graduate, employers assume that this is a fundamental requirement and a must-have skill that should not be explicitly specified in the job requirement description.

Moreover, according to our skill demand analysis, application skills in security and network are the top in-demand skills by the Saudi IT market. This is also consistent with a recent finding from (Al-Khalifa, 2017), which indicated that Network and Security are two areas that are in demand in the current market. Conversely, these skills scored below average in our IT graduates survey. There have been recent developments in skills teaching and technical learning in curriculum within local universities over the years. This development aims at providing the required skills within the labor market. There is a notable change and room for improvement, particularly within network and security applications. The second finding of this study is related with soft skills. Our analysis of market requirements (Fig. 7(b)) and senior IT and graduates survey (Table V) suggests positive results because the top in-demand skills (communication and teamwork skills) are possessed by senior IT and graduates. However, we believe that this may not be true. Students who have graduated from a university program assume that they are good communicators and strong team players. According to (Stevens & Norman, 2016), who surveyed and interviewed 12 employers in the local IT market, graduates

from the universities are unequipped with the required soft skills.

For certifications, the top in-demand certificate in the IT Saudi Market (Fig. 7(c)) is Cisco. However, only 10% of the senior and IT graduates have it as revealed by the survey results (Fig. 13). Furthermore, our findings show that only 21% of respondents hold IT certificates and 61.90% of those are male. Looking closely, it can be concluded that certifications including Cisco, ITIL, and PMP will add value to the resume of the job titles as they are high in demand in the IT Saudi labor Market.

Another interesting finding suggests Python, C#, and JavaScript as the most in-demand programming languages in the IT Saudi Market (Fig. 7(e)). However, these were pointed as major areas of weakness in the possessed skills by IT graduates according to the survey conducted (Table VI). This could be because unlike Java, Python and C# are not offered as separate courses in educational institutions. In addition, from the most in-demand programming languages, Java, CSS, and HTML are possessed by senior IT and graduates. One interesting finding in the survey is that MySQL scored high in the possessed skills by IT graduates, whereas it is not in-demand in the market.

6. Conclusion and Future Work

This study aimed to identify the in-demand skills for recruitment in the IT Saudi labor market from the employer's point of view and to allow IT graduates to predict their suitability for recruitment in the IT Saudi labor market. Finally, this study aimed to provide data-driven recommendations of the in-demand skills in IT labor market in Saudi Arabia to overcome the unemployment problem. Results showed that there existed a gap between labor market employers' expectation of Saudi workers and the skills that the workers were equipped with from their educational institutions. Planned collaboration between industry and education providers is required to narrow down this gap by providing courses that address these skills.

We cannot rely on hard skills, soft skills like communication, problem solving, and teamwork are increasingly significant. Yet, university courses and faculty should help students develop nontechnical skills. Furthermore, as IT certifications allow employers to shortlist potential candidates more quickly, educational institutions should offer academic credits for the completion of such certifications and encourage students to pursue them.

The implications of this study are beneficial for the academia to better align their educational programs with changing market requirements and improve the in-demand skills of the Saudi IT market among IT students. The study left a scope for future work to extend the survey to IT managers and faculties in IT-related academic programs. Moreover, the COVID-19 pandemic has changed the local and global labor market requirements; hence, current and future studies in this area may want to consider using data from LinkedIn's economic graph to analyze the market and investigate the implications of COVID-19 on the IT labor market of Saudi Arabia.

Acknowledgment

The authors would like to thank the Deanship of Scientific Research and RSSU at King Saud University for their technical support.

References

ACM. (2017). Task Group on Information Technology Curricula. Information Technology Curricula 2017: Curriculum Guidelines for Baccalaureate Degree Programs in Information Technology. New York, NY, USA: Association for Computing Machinery.

ACT. (2011). A better measure of skills gaps: utilizing ACT skill profile and assessment data for strategic skill research. ACT.

Aken, A., Litecky, C., Ahmad, A., & Nelson, J. (2010). Mining for Computing Jobs. *IEEE Software*, *27*(1), 78–85. https://doi.org/10.1109/MS.2009.150

Al-Khalifa, H. (2017). A survey of IT jobs in the Kingdom of Saudi Arabia 2017. Retrieved December 10, 2020, from https://www.slideshare.net/hend.alkhalifa/a-survey-of-it-jobs-in-the-kingdom-of-saudi-arabia-2017

Alsafadi, L., & Abunafesa, R. (2012). ICT Skills Gap Analysis of the Saudi Market. In *The World Congress on Engineering and Computer Science*. San Francisco, USA.

Bayt.com. (2020). Retrieved August 1, 2020, from https://www.bayt.com/

Bell, J. (2014). *Machine Learning: Hands-On for Developers and Technical Professionals*. Wiley.

Bhulai, S. (2016). Analysing which factors are of influence in predicting the employee turnover.

Boone, H. N., & Boone, D. A. (2012). Analyzing Likert data. *Journal of Extension*.

Brownlee, J. (2018). A Gentle Introduction to k-fold Cross-Validation. Retrieved December 10, 2020, from https://machinelearningmastery.com/k-fold-cross-validation/

Charitable, R. (2011). *A Model for Predicting a Career Success in Engineering Among Women and African American Men*. ProQuest Dissertations and Theses.

CITC. (2015). ICT workforce in the Kingdom of Saudi Arabia. The Communications and Information Technology Commission (CITC).

Eclipse. (2016). Eclipse. Retrieved from https://eclipse.org/ide/

El-Gabaly, M., & Majidi, M. (2003). The Information Communication Technology (ICT) Penetration and Skills Gap Analysis (SGA). Planning and Learning Inc.

Ericsson, M., & Wingkvist, A. (2014). Mining job ads to find what skills are sought after from an employers' perspective on IT graduates. In *the 2014 conference on Innovation & technology in computer science education (ITiCSE '14). Association for Computing Machinery, New York, NY, USA, 354*. https://doi.org/https://doi.org/10.1145/2591708.2602670

Gallivan, M., Truex, D., & Kvasny, L. (2002). An analysis of the changing demand patterns for information technology professionals. In *the 2002 ACM SIGCPR conference on Computer personnel research (SIGCPR '02). Association for Computing Machinery, New York,*

NY, USA, 1–13. https://doi.org/https://doi.org/10.1145/512360.512363

Hamoud, B., & Atwell, E. (2016). Quran question and answer corpus for data mining with WEKA. In *Proceedings of 2016 Conference of Basic Sciences and Engineering Studies, SGCAC*. https://doi.org/10.1109/SGCAC.2016.7458032

Huang, H., Kvasny, L., Joshi, K., Trauth, E., & Mahar, J. (2009). Synthesizing IT job skills identified in academic studies, practitioner publications and job ads. In *the special interest group on management information system's 47th annual conference on Computer personnel research (SIGMIS CPR '09). Association for Computing Machinery, New York, NY, USA, 121–128*. https://doi.org/https://doi.org/10.1145/1542130.1542154

Ibeaheem, H., Ragmoun, W., & Elawady, S. (2017). The role of Saudi Universities on the improvement of higher education skills on Saudi Arabia. *The Business and Management Review*, *9*(2).

Kennedy, H. (2019). The Labor Market in Saudi Arabia: Background, Areas of Progress, and Insights for the Future. Harvard Kennedy School.

Khan, A., Baharudin, B., Lee, L., & Khan, K. (2010). A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances In Information Technology*, *1*(1). https://doi.org/10.4304/jait.1.1.4-20

Kilhoffer, Z. (2020). Report on how to identify and compare newly emerging occupations and their skill requirements. Deliverable 12.2, Leuven, InGRID-2 project 730998 – H2020.

Krippendorff, K. (2018). *Content Analysis: An Introduction to Its Methodology* (FOURTH EDI). SAGE Publications, Inc.

Labor market reports. (2020). Jadwa Investment.

Lending, D., & Dillon, T. (2013). Identifying skills for entry-level IT consultants. In *the 2013 annual conference on Computers and people research (SIGMIS-CPR '13). Association for Computing Machinery, New York, NY, USA, 87–92*. https://doi.org/https://doi.org/10.1145/2487294.2487311

Linkedin. (2020). Retrieved August 1, 2020, from https://www.linkedin.com/

MCIT. (2014). IDC: Saudi IT Spending To Reach $14.2bn In 2017. Retrieved July 3, 2020, from available: https://www.mcit.gov.sa/en/media-center/news/92185

Mishrif, A., & Alabduljabbar, A. (2018). Quality of Education and Labour Market in Saudi Arabia. *In: Mishrif A., Al Balushi Y. (Eds) Economic Diversification in the Gulf Region*, *1*(The Political Economy of the Middle East. Palgrave Macmillan, Singapore). https://doi.org/https://doi.org/10.1007/978-981-10-5783-0_5

Octoparse. (2020). Retrieved November 10, 2020, from https://www.octoparse.com/

OpenRefine. (2020). OpenRefine Tool. Retrieved October 10, 2020, from https://openrefine.org/

Paparrizos, I., Cambazoglu, B., & Gionis, A. (2011). Machine learned job recommendation. In *the fifth ACM conference on Recommender systems (RecSys '11). Association for Computing Machinery, New York, NY, USA* (pp. 325–328). https://doi.org/https://doi.org/10.1145/2043932.2043994

Punnoose, R., & Ajit, P. (2016). Prediction of Employee Turnover in Organizations using Machine Learning Algorithms. *International Journal of Advanced Research in Artificial Intelligence(IJARAI)*, *5*(9). https://doi.org/http://dx.doi.org/10.14569/IJARAI.2016.050904

Sen, S. (2011). MANAGERIAL PERSPECTIVES OF INFORMATION TECHNOLOGY SKILLS FOR COMPUTER TRAINING INSTITUTES. *International Journal of Arts & Sciences*, *4*(12), 267–282.

SSDS. (2010). Numeracy Skills.

Stevens, M., & Norman, R. (2016). Industry expectations of soft skills in IT graduates: a regional survey. In *the Australasian Computer Science Week Multiconference (ACSW '16). Association for Computing Machinery, New York, NY, USA, Article 13, 1–9*. https://doi.org/https://doi.org/10.1145/2843043.2843068

Tapado, B., Acedo, G., & Palaoag, T. (2018). Evaluating information technology graduates employability using decision tree algorithm. In *the 9th International Conference on E-Education, E-Business, E-Management and E-Learning* (pp. 88–93).

Waikato. (2016). Weka 3 - Data Mining with Open Source Machine Learning Software in Java. *The University of Waikato*.

Wowczko, I. (2015). Skills and Vacancy Analysis with Data Mining Techniques. *Informatics*, *2*(4), 31–49. https://doi.org/https://doi.org/10.3390/informatics2040031

Zhu, C., Zhu, H., Xiong, H., Ma, C., Xie, F., Pengliang, D., & Li, P. (2018). Person-Job Fit: Adapting the Right Talent for the Right Job with Joint Representation Learning. *ACM Transactions on Management Information Systems*, *9*(3), 17 pages. https://doi.org/https://doi.org/10.1145/3234465